Development of A Novel Virtual Tool for Donor Heart Fitting

by

Jonathan Douglas Plasencia

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2018 by the
Graduate Supervisory Committee:

David Frakes, Co-Chair
Vikram Kodibagkar, Co-Chair
Rosalind Sadleir
Yiannis Kamarianakis
Steven Zangwill
Stephen Pophal

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

Heart transplantation is the final treatment option for end-stage heart failure. In the United States, 70 pediatric patients die annually on the waitlist while 800 well-functioning organs get discarded. Concern for potential size-mismatch is one source of allograft waste and high waitlist mortality. Clinicians use the donor-recipient body weight (DRBW) ratio, a standalone metric, to evaluate allograft size-match. However, this body weight metric is far removed from cardiac anatomy and neglects an individual's anatomical variations. This thesis body of work developed a novel virtual heart transplant fit assessment tool and investigated the tool's clinical utility to help clinicians safely expand patient donor pools.

The tool allowed surgeons to take an allograft reconstruction and fuse it to a patient's CT or MR medical image for virtual fit assessment. The allograft is either a reconstruction of the donor's actual heart (from CT or MR images) or an analogue from a health heart library. The analogue allograft geometry is identified from gross donor parameters using a regression model build herein. The need for the regression model is donor images may not exist or they may not become available within the time-window clinicians have to make a provisional acceptance of an offer.

The tool's assessment suggested > 20% of upper DRBW listings could have been increased at Phoenix Children's Hospital (PCH). Upper DRBW listings in the UNOS national database was statistically smaller than at PCH (p-values: < 0.001). Delayed sternal closure and surgeon perceived complication variables had an association (p-value: 0.000016) with 9 of the 11 cases that surgeons had perceived fit-related complications had delayed closures (p-value: 0.034809).

A tool to assess allograft size-match has been developed. Findings warrant future preclinical and clinical prospective studies to further assess the tool's clinical utility.

DEDICATION

To my family, for their encouragement and support to reach for the stars when others said I was reaching too far. Surviving congenital heart disease with post-operative stroke complications and diagnosed with a learning disability in my early childhood presented many challenges for my family that they never once complained about. To my mother, for countless hours of tutoring (Mom, I've come a long way from the times we sat on our porch steps in Virginia, while you tirelessly worked to teach me how to read an analog clock) – thank you for getting me here. To Ms. Mulvaney, Ms. DeWitte, and Mr. Phillips, for being part of an amazing team of middle school teachers that recognized my potential and worked to transition me into regular classes by the time I started high school.

And…

To children diagnosed with congenital heart disease and learning disabilities: do what you like, and do what you can. Realize your unique experiences provide you with insights that you can use to help make big changes in the world.

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor Dr. David Frakes for his mentorship and support throughout my college development and research career. He saw my strengths, knew where to challenge me, and worked to strengthen my weaknesses. He even brought to light both strengths and research interests that I did not realize I had. I am truly grateful to have had the experience to work under him.

I would like to thank Phoenix Children's Hospital and the many individuals at this institute who helped me throughout my research. I would like to thank Dr. Justin Ryan and the hospital's Cardiac 3D Print Lab for helping me navigate the research experience in a clinical environment, for the tools and opportunities needed to perform the work herein, and for the overall mentorship they provided. I would like to thank Drs. Steve Zangwill and Stephen Pophal for their clinical support and mentorship throughout my research time at their hospital.

I am grateful for my engineering committee members Drs. Vikram Kodibagkar and Rosland Sadleir for their feedback, insights, and/or recommendations they provided me with my thesis and overall academic career. Through their mentorship I was able to better identify artifacts in my medical data and they sparked my investigation herein to implement the concepts from the field of allometry. I would like to give a particular thank you to Dr. Yiannis Kamarianakis for helping me develop the statistical skills needed for the work herein and willingness to strengthen my committee's statistical background by joining my Ph.D. committee.

It has been an amazing experience working in Dr. Frakes' Image Processing Applications Laboratory since I was an undergraduate student. I am grateful for the mentorships I had under Drs. Christine Zwart and Haithem Babiker, they were the original, amazing Ph.D. students that were very hands on with my early research development. They were the ones that pushed me to apply for the National Science Foundation Graduate Student fellowship that I was awarded. I have to call out Dr. Priya Nair, she was an amazing labmate (turned postdoc) and friend that helped guide me throughout the Ph.D. and graduate student mazes. I would like to thank Keawepono (Pono) Wong, for his willingness to take on many of my general lab responsibilities as I began to

focus more and more of my time on completing my Ph.D. project. I would also like to thank Drs. Sudarshan Ragunathan, Rafeed Chauhury, and Aprinda Indahlastari, and also Hooman Farsani and Wei Wei. These individuals and many other labmates worked with me on various projects and/or traveled through the graduate student maze with me.

To my close friends, thank you for helping me to keep my sanity, sometimes to the expense of your own. To my family, for their love and support.

Funding sources for this research includes: Phoenix Children's Hospital Research Award Committee Grant, the National Science Foundation Award (#1512553), and the National Science Foundation Graduate Fellowship (DGE-1311230). Dr. Frakes' ASU Women and Philanthropy Society grant for "3D Printing for Heart Surgery Planning" funding also supported this work. I would like to also thank the Sharon D Lund Foundation and the Phoenix Children's Hospital Foundation for support of the Cardiac 3D Print Lab at Phoenix Children's Hospital, who provided the tools and opportunities for the work herein.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

2 PULLING FROM GENERALIZED KNOWLEDGE IN THE FIELD OF ALLOMETRY 5

2.1 Introduction to the Square-Cube Law and the Field of Allometry 5

2.2 Cardiac Relevant Allometry relationships within Mammalians 10

2.3 Developmental Growth and Overweight Effects on Scaling Relationships 13

2.4 What is the Postulated Scaling Exponent of the Mammalian Heart? 14

3 BACKGROUND AND SIGNIFICANCE FOR PROPOSED VIRTUAL FIT ASSESSMENT TOOL 15

3.1 Heart Transplant Allocation and the Allograft Shortage 16

3.2 The Clinical Standard for Allograft Size Matching and Concerns for Mismatch 19

3.3 Limitations of the DRBW Ratio Method and other Previously Proposed Solutions 21

3.4 Proposal of a Virtual Heart Transplant Fit Assessment Tool 24

3.5 Types of Heart Transplants 27

4 DEVELOPMENT OF THE HEALTHY HEART LIBRARY 31

4.1 Methods and Materials of the Healthy Heart Library 32

4.2 Results of the Healthy Heart Library 36

4.3 Discussion of the Healthy Heart Library 44

5 DEVELOPMENT OF THE INITIAL AND IMPROVED ALLOGRAFT TCV PREDICTION MODELS 53

5.1 Methods and Materials of the Allograft TCV prediction Model 54

5.2 Results of the Allograft TCV prediction Model 64

5.3 Discussion of the Allograft TCV prediction Model 104

5.4 Methods and Materials of the Improved Allograft TCV prediction Models 113

CHAPTER Page

5.5 Results of the Improved Allograft TCV prediction Models 114

5.6 Discussion of the Improved Allograft TCV prediction Models 125

6 ANALYSIS OF THE VIRTUAL HEART TRANSPLANT FIT ASSESSMENT TOOL'S CLINICAL UTILITY 129

6.1 Methods and Materials for Virtual Heart Transplant Fit Assessments 130

6.2 Methods and Materials for Study 1: Expanding Upper Donor Pool Ranges 130

6.3 Methods and Materials for Study 2: Perceiving Fit-Related Complications 131

6.4 Results for Study 1: Expanding Upper Donor Pool Ranges 132

6.5 Results for Study 2: Perceiving Fit-Related Complications 141

6.6 Discussion of Tool's Clinical Utility 143

6.7 Discussion on Virtual Transplant Case Study Using Donor Images and the Tool 147

7 LIMITATIONS, FUTURE WORK, AND CONCLUSIONS 150

7.1 Limitations of the Healthy Heart Library and Allograft TCV Prediction Model 151

7.2 Limitations of the Clinical Utility Assessment 152

7.3 Future Work 155

7.4 Conclusions 158

REFERENCES 160

APPENDIX

A LIST OF ACRONYMS 170

B EFFECT SIZE AND SAMPLE SIZE NEEDED FOR TWO MEAN COMPARISON 173

C DELEGATION OF TCV MODEL DEVELOPMENT AND VALIDATION 177

D COPYRIGHT PERMISSIONS 179

LIST OF TABLES

Table		Page
2.1	Empirical Scaling Relationships for Metabolism	11
3.1	Waitlist Mortality for Pediatric Weight Groups	16
4.1	Healthy Heart Library Data	33
4.2	Demographics of the Heathy Heart Library (N=97)	37
4.3	Allometric Scaling Signals of Geometric Metrics with TCV: With Influential Outliers	42
4.4	Allometric Scaling Signals of Geometric Metrics with TCV: Without Influential Outliers	43
4.5	Allometric Exponent Scaling Signals of Geometric Metrics with TCV	44
4.6	Healthy Heart Library Pearson's Correlations	44
5.1	Considered Regression Modeling Variables	55
5.2	The Formulas used to Quantify the Statistical Metrics	57
5.3	Potential Models Identified by AICc Difference Criteria	68
5.4	Potential Models' Cross-Validated Mean Abolute Errors	71
5.5	Mean Absolute Errors: Summary and Comparison	72
5.6	Potential Models' Cross-Validated Mean Abolute Percentage Errors	73
5.7	Mean Absolute Percentage Errors: Summary and Comparison	74
5.8	Potential Models' Cross-Validated Mean Errors	75
5.9	Mean Errors: Summary and Comparison	76
5.10	Potential Models' Cross-Validated Mean Percentage Errors	77
5.11	Mean Percentage Errors: Summary and Comparison	78
5.12	Potential Models' Cross-Validated Mean Square Errors	79
5.13	Mean Square Errors: Summary and Comparison	80
5.14	Potential Models' Cross-Validated Root Mean Square Errors	81
5.15	Root Mean Square Errors: Summary and Comparison	82
5.16	Testing for Heteroscedasiticy	83
5.17	Heteroscedasticity Correction Results	85
5.18	Final Results for Initial Modeling Process: Model 1	88

Table Page

5.19 Final Results for Initial Modeling Process: Model 2 89

5.20 Final Results for Initial Modeling Process: Model 3 90

5.21 Final Results for Initial Modeling Process: Model 4 91

5.22 Final Results for Initial Modeling Process: Model 5 92

5.23 Final Results for Initial Modeling Process: Model 6 93

5.24 Final Results for Initial Modeling Process: Model 7 94

5.25 Final Results for Initial Modeling Process: Model 8 95

5.26 Final Results for Reference Model With Lowest AICc Rank 96

5.27 Final ln(Ht) Coefficent Comparison for the 8 Top AICc Models 97

5.28 Final Testing Statistical Model Metric Results 100

5.29 Final Ranked Testing Statistical Model Metric Results 100

5.30 Final Ranked Testing Statistical Model Metric Results with sMAPE and sMPE 101

5.31 Final Results for Improved Models' B and B* 116

5.32 Testing Statistcal Metric Results for Initial and Improved Models 117

6.1 PHTS, UNOS, and Virutal Maximum DRBW Ratios 134

6.2 PCH and Virtual Maximum DRBW Ratios 135

6.3 PHTS, UNOS, and Virtual Maximum DRBW Ratios 136

6.4 Accepted Allograft Size-Match Percentage by Classifciation 137

6.5 Approximate Sample Sizes Needed to Detect DRBW Ratio Differences 138

6.6 Frequency and Type of Surgeon Percieved Fit-Related Complications 142

6.7 Delayed Sternal Closure vs. Surgeon Perceived Fit-Related Complication Matrix 143

LIST OF FIGURES

Figure		Page
4.1	Healthy Heart Reconstruction Showing TCV Boundry	35
4.2	Plot of Healthy Heart Library Deomographics (N=97)	38
4.3	Color-Coded Plot of Healthy Heart Library Deomographics (N=97)	39
4.4	Healthy Hearts: Age vs BMI	40
4.5	mTCV by Modality	41
5.1	The Key Steps of the Initial Modeling Process	56
5.2	AICc Profiles from Initial Modeling Process's Exhastive Search	67
5.3	Model A's pTCV vs. the mTCV	102
5.4	Relative Error between Model A's pTCV and the mTCV	103
5.5	Normal Q-Q Plot for Model A's pTCV Residuals	103
5.6	Histogram for Model A's pTCV Residuals	104
5.7	Examples of Interpolation and Extrapolation Predictions	110
5.8	Model A, B, and B*'s pTCV vs. mTCV	118
5.9	Relative Error between Model A, B, B*'s pTCV and the mTCV	119
5.10	pTCVs of Model A vs. B	120
5.11	Relative Error between Model A and B's pTCV	121
5.12	Model A and B's pTCV vs. the mTCV: Zoomed In	122
5.13	Relative Error between Model A and B's pTCV and the mTCV: Zoomed In	123
5.14	Normal Q-Q Plot for Model B's pTCV Residuals	124
5.15	Histogram for Model B's pTCV Residuals	124
6.1	Phoenix Children's Hospital and Virtual Maximum DRBW Ratios	135
6.2	PHTS, UNOS, and Virtual Maximum DRBW Ratios	136
6.3	Localized Smoothing Spline for DRBW Ratios	140
6.4	Cases Tool Suggested PCH's Upper Listing could be Expanded	141
6.5	Virtual Fit Assessment for Case Study	148
B.1	Sample Size Vs. Effect Size for Two Mean Comparison	176

Figure Page

D.1 Permissions For: Figure 6.3 180

CHAPTER 1

INTRODUCTION

The heart transplant (HTx) is the final treatment option for end-stage congestive heart failure. A clinically-concerning problem for patients needing a HTx in the United States of America (USA) is the current donor-shortage. Recent publications found that approximately 1 in 10 adults die while waiting for a HTx [1–4]. The waitlist mortality rate is higher in pediatrics due to the limited number of appropriately sized organs, i.e., allografts. Mechanical treatment options for heart failure are traditionally stop-gap measures for either (1) bridge-to-recovery, (2) bridge-to-destination, or (3) bridge-to-transplant. Furthermore, mechanical treatment options may not be compatible with pediatric patients and even smaller adult sizes due to overall implantable device volume. The unfortunate reality in this donor-shortage is 70 pediatric patients die annually while 800 healthy, well-functioning cardiac allografts get discarded [5,6]. The donor-recipient body weight (DRBW) ratio is the clinical standard to evaluate allograft size-match but neglects anatomical aberrations and patient-to-patient differences that are within the norm. Concerns of potential size-mismatch is one source of allograft waste and waitlist mortality.

This dissertation presents the development and validation of a virtual HTx fit assessment tool to assess allograft fits. This tool relies on a library of healthy hearts and a regression model to predict an allograft's total cardiac volume (TCV). Specifically, the development and validation of the library and regression model will be presented. The clinical utility of the tool will be investigated in a series of initial, retrospective, clinical studies.

The hypothesis of this research is volumetric medical images, i.e., computed tomography (CT) or magnetic resonance (MR) images, can be leveraged to provide additional clinical information related to HTx allograft size-matching of an offered donor. A novel virtual HTx fit assessment tool will be developed and validated to help leverage medical images for fit assessment. It is further hypothesized an early investigation into the tool's clinical utility will produce

trends that suggest the tool can help safely expand patients' donor pools and therefore improve allograft size-matching.

The author's, i.e., Plasencia's, pivotal contributions to the fields of biomedical engineering and medicine will be:

1. a quantitative understanding of the biological signal relationships between the healthy TCV and various gross subject parameters within the pediatric and young adult population,

2. the development and validation of a regression model that predicts normal, healthy heart TCVs while accounting for pediatric and young adult developmental growth and body type,

3. the demonstration and the initial validation into the tool's relevance in allograft size-matching, and

4. an initial investigation into whether clinicians can perceive allograft fit-related complications when fusing reconstructions onto a medical image.

The author hopes this work further contributes to medicine by initiating other groups to investigate the strengths and short comings of the novel tool presented herein. The author hopes these groups will develop and validate their own improved virtual fit assessment methods for the clinical environment in which HTx and other areas of solid-organ transplant are considered.

Chapters 2 and 3 are background and significance chapters. Chapter 2 introduces fundamental biological scaling relationships from the field of allometry that the allograft TCV prediction modeling process exploited. Chapter 3 introduces an overview of HTx and waitlist statistics and procedures. Furthermore, chapter 3 discusses current allograft size-matching

methods and its limitation to aggressively expand patient donor pools so clinicians can procure an allograft in-time for their patient.

Chapters 4, and 5 cover tool development and prediction model validation. Chapter 4 develops a library of normal, healthy hearts with patient parameters and TCV reconstructions included. Chapter 5 uses (1) the library, (2) linear regression techniques, and (3) concepts from the field of allometry to develop and validate an allograft TCV prediction model. The validation procedure in chapter 5 looks at the testing error of the developed prediction model but not the clinical utility.

Chapter 6 covers early retrospective study validation of the tool's clinical utility in 2 main outcome analysis studies. The first study investigates if the tool can alter surgeon perception to accept allografts that are traditionally considered oversized and therefore aid clinicians to expand donor pools. The second study investigates if the tool allows clinicians to perceive fits that are safe and fits that are not safe. The chapter concludes with the presentation of an additional, unique case study in which donor images were available for virtual assessment.

Finally, chapter 7 covers the capital successes, limitations, future works, and conclusions of this novel tool. This includes an analysis of both the assumptions that had to be made in this body of work and the comprises needed to be made because (1) this is human research and (2) this is a retrospective, chart review study.

The work presented herein are covered by the Phoenix Children's Hospital (PCH) institutional review board (IRB) protocols #16-131 and #17-031. The additional case study was previously presented by the author as a case study and therefore was determined by the PCH IRB to be non-human research.

Lastly, the author wants to bring to the reader's attention that it is commonplace for the clinical environment to report body weight in a unit of "kilogram", i.e., "kg". The author recognizes "pound" is a unit of weight and "kg" is a unit of mass from different scientific unit systems. Multiplying "kg" by the gravitational acceleration constant produces a weight value. The author has taken it upon himself to intentionally not distinguish between body weight and body mass herein, unless stated otherwise, to ease confusion by using the established clinical terminology and reporting

practices. The common clinical practices related to this body of work are (1) “weight” (Wt) is often reported in kgs and (2) the DRBW ratio interchangeably relates donor and recipient body masses (using kgs) and body weights (using lbs). Clinicians can interchangeably use either kgs or lbs for the DRBW ratio – assuming the numerator and denominator keep the same unit – because of the linear relationship between the two units.

CHAPTER 2

PULLING FROM GENERALIZED KNOWLEDGE IN THE FIELD OF ALLOMETRY

The novel virtual HTx fit assessment tool developed within this body of work requires a regression model to predict allograft TCV from gross donor parameters. The prediction model is required because measuring an allograft TCV will, in general, be impractical for the majority of clinical scenarios. To develop the regression model, chapter 2 will cover the field of allometry and consider if the knowledge from this field can potentially be implemented to predict allograft TCVs. The field of allometry is a subset of biology in which biologist mathematically study biological scaling relationships [7].

The review within this chapter will (1) provide an overview of the field of allometry and (2) introduce the reader to the power-law functions the field deploys in size scaling analysis. In the first half of this chapter the reader will be introduced to the initial contributions Galileo Galilei had on the field of allometry, understand the mathematical strength and simplicity of the power-law functions in biological scaling, and be able to identify isometric versus allometric scaling relationships. In the second half of this chapter, a simple, retrospective analysis on previous respiratory and cardiovascular empirical results will be performed to demonstrate the relevance of the power-law functions in studying the cardiovascular system. Additionally, the author will postulate what scaling relationship between TCV and body mass is to be expected. The consequences of the postulated scaling relationship will be discussed in the following chapter 3. At the end of chapter 2 the reader should, at a minimum, understand why the author considered it worthwhile to investigate if the power-law functions could be used to predict an allograft's TCV.

2.1 Introduction to the Square-Cube Law and the Field of Allometry

Dialogues Concerning Two New Sciences (Galileo Galilei in 1638) is traditionally attributed as the original publication that described a physical phenomenon that is now known as the square-cube

law [8]. Galileo observed that the ratio of volumes of two similar geometries will always be greater than the ratio between their corresponding surfaces. Galileo's "greater than" observation assumes the larger object is placed in the numerator position. This observation introduces, in an indirect manner, that a volume of an object increases at a faster rate than its corresponding surface. It is now understood this phenomenon is a consequence of volume having its characteristic unit, i.e., length, cubed while area has its characteristic unit squared. In other words, this a consequence of volume and area being reported in m^3 and m^2, respectively.

The phenomena described by the square-cube law is by no means isolated to volumes and areas. It could be another spatial relationship such as volume and length, or it can be a completely different characteristic unit relationship like time is with velocity and acceleration. The early geometric ratio shape analysis Galileo performed is now often analyzed by using power-law functions to scale both geometric and non-geometric relationships.

The simplicity and berth of the power-law functions to scale quantities have been the pillars for the adoption of this mathematical tool in many fields of modern science, engineering, and manufacturing [8–19]. Applying power-law functions to anatomical, physiological, and other biological relationships is fundamental to the field of allometry [7,16,20,21]. Within biology and medicine, this scaling methodology has been used in various studies and applications that include but are not limited to the investigation of biological size limits, prediction of size-based biological needs, understanding of size-based toxin and drug clear rates, scaling of pharmaceutical trial study dosages from animal to human, and reduction of adult drug dosages for pediatrics [7,22–30].

The generic power-law function, Equation 2.1, scales the relationship between the two quantities "*x*" and "*y*" in which "*a*" is the proportional constant and "*b*" is the scaling exponent:

$$y = a * x^b \qquad \text{(Equation 2.1)}$$

The proportional constant, "*a*", quantitatively describes "how" the two quantities are different in Equation 2.1. This proportional constant is independent of size scaling. As its name implies, the scaling exponent, "*b*", contains the scaling information but does not describe "how" the quantities

are different outside of scaling. To better visualize the meaning of the proportion constant and scaling exponent, the reader can think about a basic geometric shape like the sphere.

The governing equation for the volume of a sphere is $Volume = \frac{4}{3} * \pi * radius^3$. Immediately it should be obvious that this geometrically-descriptive equation is a power-law function (i.e., Equation 2.1) in which $a = \frac{4}{3} * \pi$ and $b = 3$. If the reader visualizes the volume and radius of a sphere as two different quantities, i.e., two different objects, that are dependent on one another then the reader can view the constant "*a*" as a quantitative way to describe "how" these two "geometries" relate. The descriptive nature of "*a*" can be better appreciated when comparing the sphere's volumetric equation to the equations for (1) a cube's volume, (2) a regular tetrahedron's volume, (3) a regular octahedron's volume, or (4) even between the volume and surface area equations of a sphere with radius as the input. This "how" just happens to be a scalar value description. It is because "*a*" describes "how" these two quantities relate (independent of size) that this variable is referred to as the "proportional constant". This "how" description by the variable "*a*" holds for non-geometric relationships. The scalar value of constant "*b*" in this example is 3 in which we are scaling the radius, i.e., a length, "*x*", to the sphere's volume, "*y*". This scaling value is expected because a volume increases by a power of 3 while the radius increases by a power of 1.

At initial glance Equation 2.1 may not appear simple; power-law functions appear to be particularly problematic when there is more than one input quantity. However, Equation 1 can be simplified when it is linearized by applying a log-log transform as seen in Equation 2.2:

$$y = a * x^b \leftrightarrow ln(y) = ln(a) + b * ln(x) \quad \text{(Equation 2.2)}$$

Similarly, a power function with two input quantities can be simplified when it is log-log transformed as seen in Equation 2.3:

$$y = a * x_1^{b_1} * x_2^{b_2} \leftrightarrow ln(y) = ln(a) + b_1 * ln(x_1) + b_2 * ln(x_2) \quad \text{(Equation 2.3)}$$

It should be noted with Equation 2.3 that the proportional constants between "y" and the inputs "x_1" and "x_2" are combined into a single proportional constant variable "*a*". The scaling exponent variables are not combined.

Power-laws are a mathematically powerful class of functions because, as Equations 2.2 and 2.3 demonstrate, the nonlinear relationships can quickly be simplified when applying a log-log transform to linearize the relationships. Once the power-law relationship is log-log transform, then the proportional constant serves as the intercept (i.e., in the form of $ln(a)$) and the scaling exponents serve as slopes that scale the relationships linearly. It is while a power-law function is linearized one could then apply generalized linear regression techniques to fit empirical data and therefore determine the proportional and scaling exponent coefficients. Using *log* based *e*, i.e., *natural log*, in the log-log transform is preferable because it allows one to take the *exp()* function to the left- and right-hand-sides of the fitted model to easily reverse the forward transform. In other words, applying the *exp()* function to the left- and right-hand-sides of the fitted model is to take the backward transform. Reporting the fitted model after the backwards transform is applied allows the input(s) and output to be trivial because these variables no longer need to be log transformed before being inputted into the developed model.

Studies within the field of allometry will often fit empirical data of a given power-law function to either interpret the biological significance or make biological predictions [7]. A fundamental interpretation a biologist can make from the quantitative scaling exponent metric is whether a scaling relationship is isometric or allometric. Referencing the Equation 2.1 power function, an isometric scaling relationship is said to exist when the empirical ($\hat{b}$) and theoretical (b) scaling exponents are equivalent; otherwise the relationship is allometric. The theoretical scaling exponent is based on the powers of the characteristic units. For example, if the inputs "*x*" and "y" in Equation 2.1 are an area (Area=Length^2) and a volume (Volume=Length^3), respectively, then an isometric relationship postulates $\hat{b} \approx b \stackrel{\text{def}}{=} 3/2 = 1.5$.

The power-law function relationships analyzed heretofore made comparisons based on two or more quantities with the same characteristic unit. It is preferable, in general, for the quantities

being compared to have a commonality in their relationship which is to say they share a characteristic unit. Sharing a characteristic unit eases the interpretation of the scaling between the quantities and allows one to either postulate relationships before empirical data is collected or to postulate relationships that cannot be measured. However, the power-law functions do not require the inputs and outputs to have the same characteristic unit. One can either (1) mathematically relate the different characteristic units based on the use of appropriate constants or (2) fit the appropriate power-law function between the inputs and output without relating the different characteristic units [7].

Mathematically relating different characteristic units, i.e., option (1), might be preferable to option (2) for two important reasons. First, option (1) may help to simplify the interpretation of the empirical data. Second, option (1) may help to postulate relationships in which one of the quantities has yet to be measured or cannot be measured. For example, let us assume one wants to know the scaling relationship between the radius of a sphere, "*x*", and the volume of a sphere, "*y*", but rather then collecting the sphere's volume, the sphere's mass was collected instead. If the density of the sphere can be assumed to be a uniform constant then there will be a linear relationship between volume and mass. Quantitatively the scaling exponent of a linear relationship is unity, i.e., 1. Given the scaling exponent is 1 and volume has its characteristic unit (i.e., length) to the power of 3 then an appropriate power can be found to mathematically describe the scaling exponent of mass as a length; i.e., $b \stackrel{\text{def}}{\cong} 1 = L_{vol}/L_{mass} = 3/L_{mass} \rightarrow L_{mass} = 3$ with "L" being the characteristic power. This result implies, in effect, that the scaling of this sphere's mass can mathematically be described as the characteristic unit, i.e., length, to the power of 3.

The particularly usefulness of this mathematical observation for the field of allometry, i.e., option (1), is often seen when one is comparing a geometric scaling relationship to body mass – an important and easily obtainable biometric. Substituting mass for volume allows one to quickly postulate if there is an isometric or allometric scaling relationship between mass and the geometric quantity [7]. This concept will be implemented shortly to postulate the scaling exponent relationship between TCV and body mass.

2.2 Cardiac Relevant Allometry relationships within Mammalians

Allometry has been an important field of study in biology to postulate and explain the scaling of biological structural sizes and physiological processes. Aerobic metabolism scaling relationships can explain many of the scaling relationships seen in mammalian cardiovascular and respiratory systems.

The current chapter section will now retrospectively analyze historical aerobic metabolic data to support the consideration of using a power-law function to predict allograft TCV in this body of work. Relevant, historical data to be used in this analysis are scaling exponent values that were compiled in Schmidt-Nielsen's textbook, "Scaling: Why is animal size so important?", and are provided in Table 2.1 [7]. Cardiac stroke volume was not available in Schmidt-Nielsen's textbook, therefore, it was acquired from *West et al.*'s journal publication [31].

Mammals consume oxygen to perform aerobic metabolism – the major metabolic process mammals can sustain long-term to release the energy needed for cellular function [32–34]. The scaling exponent value between oxygen consumption and mammalian body mass is 0.76. Given the lungs and heart have key functions in transporting oxygen for aerobic metabolism it could be postulated that the scaling relationships of these biological systems should relate to the scaling of oxygen consumption.

Empirical Scaling Relationships for Metabolism

Respiratory		**Unit**	$\widehat{b}$
	Lung Volume	ml	1.06
	Tidal Volume	ml	1.04
	Breath Rate	1/min	-0.26
	Ventilation Rate	ml / min	0.80
Cardiovascular			
	Heart Mass	kg	0.98
	Stroke Volume	ml	1.03
	Heart Rate	1/min	-0.25
	Cardiac Output	ml / min	0.81
Metabolism			
	Oxygen Consumption	ml / min	0.76

Table 2.1: The presented scaling relationships are with respect to body mass. The proportional constant does not drive scaling relationships and therefore has been excluded. Compiled data is from Schmidt-Nielsen's textbook, "Scaling: Why is animal size so important?" and *West et al.*'s publication [7,31].

The amount of oxygen delivered to the body for aerobic metabolism is driven by the ventilation of the lungs and the cardiac output of the heart. Assuming the efficiency of oxygen consumption is good but not 100%, then one could postulate the scaling exponents for ventilation and cardiac output might be slightly larger than that of oxygen consumption. It is well known and logical that chemical reactions, in general, will never have a 100% yield efficiency [35]. The empirical ventilation and cardiac output rates presented in Table 2.1 (0.80 and 0.81, respectively), in fact, do have a slightly greater exponent scaling value. The field of allometry can further analyze the anatomical and physiological scaling relationships of the respiratory and cardiovascular systems by considering their classical governing equations.

The classical governing equations for ventilation of the lungs and cardiac output of the heart are as follows [33,34]:

$$Ventilation \stackrel{\text{def}}{=} Tidal\ Volume * Breath\ Rate \qquad \text{(Equation 2.4)}$$

$$Cardiac\ Output \stackrel{\text{def}}{=} Stroke\ Volume * Heart\ Rate \qquad \text{(Equation 2.5)}$$

Given we are only interested in scaling relationships, proportional constants will not be considered, i.e., they were excluded in Table 2.1. The justification for excluding the proportional constants is they do not contain scaling information. In implementing the above assumption, the provided empirical exponential scaling factors from Table 2.1 can now be plugged into Equations 2.4 and 2.5 to investigate if the scaling relationships hold. The scaling relationships are:

$$Mass^{0.80} \cong Mass^{1.04} * Mass^{-0.26}$$

$$0.80 * \log(Mass) \cong 1.04 * \log(Mass) - 0.26 * \log(Mass)$$

$$0.80 \cong 0.78,$$

for the respiratory system and:

$$Mass^{0.81} \cong Mass^{1.03} * Mass^{-0.25}$$

$$0.81 * \log(Mass) \cong 1.03 * \log(Mass) - 0.25 * \log(Mass)$$

$$0.81 \cong 0.78,$$

for the circulatory system. The scaling relationships are determined to be approximately equivalent and mathematically imply that tidal volume, breath rate, stroke volume, and heart rate must be balanced for a mammal to sustain an adequate oxygen consumption rate for aerobic metabolism. These mathematical cardiovascular and respiratory relationships with oxygen consumption are expected because (1) these biological systems are responsible for oxygen delivery and (2) aerobic metabolism is the only metabolic process a mammal can sustain long-term for cellular function [33,34].

The mathematical implications suggest cardiovascular and other organ systems scale to meet the biological needs of healthy subjects. In general, this analysis of aerobic metabolism supports the stance that concepts from the field of allometry can describe how organ systems scale

mathematically by using power-law functions. These results provided the author with enough evidence that the power-law functions should be considered as a potential method to predict allograft TCV.

2.3 Developmental Growth and Overweight Effects on Scaling Relationships

A limitation with the field of allometry is that the majority of historical studies do not consider developmental growth. Many of the well-established metabolic and cardiovascular relationships in the field of allometry are developed using either data from various adult mammal species or data from within a single, adult mammalian species [7,23,24,36]. In fact, the analysis in the previous chapter section used only adult mammalian derived data.

The works of de Simone *et al.* and Bide *et al.* strongly suggest, in general, that the well-established adult scaling relationships that are typically referenced from the field of allometry do not capture developmental growth [23,37]. The work of de Simone *et al.* specifically demonstrated human cardiac scaling relationships between adults and pediatrics do not hold [37]. These findings emphasized the importance that particular caution should be taken when using historical scaling relationships from the field of allometry to compare adults and youth.

A second limitation with the field of allometry is that the majority of historical studies do not consider overweight. However, relatively recent clinical publications have been implementing the power-law functions to either compare or normalize the scaling exponent values of cardiac metrics between healthy and overweight individuals [37–41].

The implications of these key limitations are (1) children do not scale downwards as small adults and (2) body type (e.g., overweight) affect the scaling relationships. To effectively predict allograft TCV of realistic donors for a children's transplant center, which will include pediatric donors, a model needs to exist that considers both developmental growth and body type. Capturing a wide range of subject time points in the developmental process and a wide range of body types to fit to a power-law function (or any other type of model framework) will likely be necessary to realistically model the allograft TCV.

2.4 What is the Postulated Scaling Exponent of the Mammalian Heart?

Before ending this chapter, it will be useful to postulate what the scaling exponent is between the TCV and body mass of the adult mammal. The scaling exponent will need to be postulated because, although the scaling relationship between total cardiac mass (TCM) and body mass is well-established for the adult mammal, i.e., $\hat{b} = 0.98 \cong 1$, the author is unaware of a published scaling relationship between TCV and body mass in adult mammals [7]. Having a postulation of what the scaling exponent value is for adult mammals will be useful to present a potential limitation of the DRBW ratio as a standalone clinical metric for allograft size-fit assessments in chapter 3. This potential limitation will be one argument to support the need for a better allograft size-matching method in pediatric transplants at a minimum.

There are three requirements to postulate what the scaling exponent between TCV and body mass might be. The postulation argument is based on an earlier example from this current chapter in which the density of a sphere was assumed constant. First, various adult mammalian studies have shown that body weight and TCM have a 1:1 scaling, i.e., body weight and TCM scale isometrically as $\hat{b} = 0.98 \cong 1$ implies [7,20,21,27]. Second, it is a given that cardiac tissue has a constant density [42–45]. Third, it must be assumed that the blood volume within the TCV scales isometrically with TCM. If these three arguments hold then it would mathematically imply there is a 1:1:1 scaling of body weight, TCM, and TCV. Recent literature found a linear, i.e., isometric, relationship between TCV and left ventricular volume in healthy hearts [46]. If it can be assumed the remaining three cardiac chambers also scale isometrically with TCV then this publication further supports the postulation that TCV and body mass scale isometrically. This postulation that TCV and body mass scale isometrically will suggest in chapter 3 that there might be clinical implications of using adult DRBW ratio criteria in pediatrics.

CHAPTER 3

BACKGROUND AND SIGNIFICANCE FOR PROPOSED VIRTUAL FIT ASSESSMENT TOOL

Phoenix Children's Hospital clinicians initiated the collaboration that led to the development of the novel virtual HTx fit assessment tool presented herein. The clinicians seeking a better methodology to assess donor allograft fits into heterotaxy patients needing transplant. Heterotaxy is a particularly challenging group of congenital heart disease patients to find an appropriate donor fit for because many of the import anatomical structures are removed from their typical placement area, e.g., mirrored across the coronal plane. Although heterotaxy patients have extreme anatomical aberrations, the clinical standard remains the use of the DRBW ratio as a standalone metric to assess if a donor allograft will fit a patient. The metric assumes that a donor and patient have similar anatomical configurations, and weight comparisons are indicative of allograft size matching. The metric's first assumption makes the identification of an appropriately sized donor allograft particularly challenging for the heterotaxy population. Pediatric HTx fit assessments are challenging, in general, because a large subset are structural congenital heart disease cases. Donor organ shortages in the USA add additional complications to this clinical situation since doctors do not want to unnecessarily decline a donor offer for concern they may not be able to procure another organ for their patient.

Chapter 3 will (1) provide an overview of the current clinical challenges in procuring a donor allograft and (2) introduce a potential solution that will be developed to address current clinical allograft size-matching limitations. In the first half of this chapter the reader will be introduced to the current donor allograft shortage in the USA, the reader will understand key size-related considerations clinicians need to consider for fit assessment, and the reader will understand the limitations of size-fit assessment tools currently available. In the second half of this chapter the reader will be introduced to the author's proposed solution to the clinical limitations in the current allograft size-matching methodologies. The chapter will conclude with what, if any, considerations

need to be made so the proposed solution can be implemented for any of the 4 distinct types of HTx procedures.

3.1 Heart Transplant Allocation and the Allograft Shortage

Heart transplantation is the preferred treatment option in late-stage heart failure; although, donor organs (allografts) remain limited [6,47]. Pediatric patients listed for HTx (i.e., potential allograft recipients) have the highest waitlist mortality rate for any patient population needing a solid-organ donor [6]. However, half of all cardiac allografts in the USA are left unused in which 36% of the unused allografts were well-functioning allografts, i.e., 18% of all available cardiac allografts are discarded even though they are well-functioning o›rgans [5]. In practical terms, this 18% corresponds to 800 well-functioning allografts being discarded while 70 pediatric patients die on the waitlist annually in the USA [5,6]. Under the current allograft allocation policies, the USA pediatric waitlist mortality is largest for the youngest of patients, as is shown with the compiled data in Table 3.1 from *Almond et al.* [6].

Waitlist Mortality for Pediatric Weight Groups

Weight Group	*The Percentage of Waitlisted Pediatrics within a Weight Group that go to Mortality:*	*The Percentage of All Waitlist Pediatric Mortalities Represented by Weight Group:*	*The Frequency of Pediatric Waitlist Mortalities within a Weight Group:*
≤ 5kg	24%	25%	175
≤ 20kg	20%	64%	370
> 20kg	12%	36%	160
> 65kg	11%	5%	10
Total			530

Table 3.1: From *Almond et al.'s* publication, it was approximated that 24% of the smallest pediatric patients die on the waitlist while accounting for 25% of all pediatric waitlist mortalities. Although *Almond et al.* reported 533 pediatric mortalities, over a 7.5-year period, starting in 1999, the results in the current table were approximated to be 530 from a bar graph. *Almond et al.'s* results account for approximately 70 pediatric mortalities while on the waitlist annually.

The approximate 25 deaths in a group of 100 of the youngest children (by weight in Table 3.1) versus the approximate 10 deaths in a group of 100 for the largest of children demonstrates age-specific waitlist mortality discrepancies [6]. The discrepancy that the youngest of patients die on the waitlist is due to the current shortage of size appropriate allografts [6]. The mortality rate of 10 in 100 for the largest children is comparable to the adult population and therefore makes this mortality rate “acceptable” given the current overall donor shortage [1–4]. It is worth noting that critiques of the current nationwide donor shortage in the USA fault the opt-in policy while (1) Americans are in favor of organ donation and (2) opt-out European countries have larger donor rates [48]. Ultimately, the waitlist mortality rate in the young pediatric population is unacceptably high in the USA [6,49].

Allograft allocation within the USA is performed by the United Network of Organ Sharing (UNOS) organization (information can be found on their website: www.unos.org) [50]. UNOS has broken the USA into 11 distinct territories for donor allocation [50]. A patient needs to be listed in the UNOS database to receive a donor offer by the organization. To be listed in the database, a patient’s clinical team must provide clinically pertinent donor requirements and patient information. Required data includes the patient’s severity of congestive heart failure (starting with the sickest the levels are 1A, 1B, and 2) and donor requirements (e.g., acceptable weight range, blood type, etc.). Higher status patients get priority over lower status patients and longer listed patients get priority to more recently listed patients in being offered a donor. While listed, patient status may change to a 7A listing because they are either too sick or appear to be getting well enough to justify delaying a transplant in the immediate future. The intent of the 7A status is to pause the patient’s position on the waitlist while the clinical team waits to determine if, and when it is clinically appropriate to move forward with a transplant.

At donor availability, an allograft will be sequentially offered with priority going to the patient with the highest status and duration on the waitlist in which the donor matches both the listed patient’s clinical requirements and UNOS territory. If no centers accept the allograft for their patient, and if time permits it, then the allograft may be offered outside the donor’s territory with the understanding there will likely be an extended “cold ischemic time”. “Cold ischemic time” is a

clinically important time measurement recognizing that a harvested allograft, before transplantation is complete, is slowly dying while being transported. “Ischemic” refers to the lack of blood flow to the organ which causes tissue deterioration of the allograft. “Cold” refers to the organ being preserved on ice to minimize tissue deterioration during transportation. This time duration begins once the donor’s aorta is cut and continues until the allograft is attached to the recipient’s aorta. Failure to find a recipient for an allograft is frustrating for transplant teams because an unused organ is a wasted organ during the current nationwide donor shortage. Finding allografts for the very young pediatric patients are particularly challenging because many donors are outside their UNOS listed acceptance range.

To address poor pediatric outcomes UNOS implemented the current, 3-tier, “sicker first” policy, in 1999, in which pediatric patients (< 18 years) received priority allocation of adolescent allografts [6,51]. The policy that pediatric allografts are first offered to pediatric patients was more of a concern for allograft coronary artery disease associated with older donor hearts then that of the donor shortage [6,51]. The same policy reprioritized all patients < 6 months of age from top tier priority for receiving an allograft to middle tier [51]. As previously shown, Almond *et al.* demonstrated this policy yields an unacceptably high waitlist mortality rate for pediatric patients with body weights ≤ 10kg and still unfavorable rates for patients ≤ 20kg [6]. The UNOS policy likely favors older pediatric patients because these patients get priority over adults in receiving pediatric allografts and, unlike younger children, are more likely to be able to accept adult allografts while infants are forced to share allografts with older pediatric patients.

Clinicians of pediatric patients – especially when locating allografts for the very young – must expand their patient’s typical donor criteria to increase the likelihood of procuring an allograft in time given the current UNOS policies. Typical methods clinicians use to expand their recipient’s donor pool are as follows: (1) ABO-incompatible transplantations in the very young and (2) expanding listed DRBW ratio criterion [6,52].

ABO-compatible transplantations are generally required otherwise aggressive allograft rejection will occur due to an immune system response. The clinical practice of accepting ABO-incompatible allografts for infants is based on observations that the immune systems of the very

young are not yet fully developed and are unlikely to cause aggressive allograft rejection later in life if the transplant is done early enough [52]. Performing ABO-incompatible transplants allows the donor pools of infants to be safely expanded.

The DRBW ratio criterion is the standalone metric HTx centers use to assess if a donor allograft will safely fit. Expanding the lower and upper DRBW ratio criteria is seen as another method to expand a patient's donor pool. Through experience, in general, clinicians will cautiously expand the metric criteria range to expand their patient's donor pool, as is suggested in literature [53–56]. However, the DRBW ratio ranges are limited in considering unique patient aberrations typically seen in structural congenital heart disease. This limitation of the DRBW ratio metric makes it particularly challenging for clinicians to safely expand the donor pool of congenital heart disease patients based on weight. This body of work attempts to reduce allograft waste (e.g., the annual discarding of approximately 800 well-functioning hearts) and to reduce waitlist mortality through the development a novel tool that helps clinicians safely expand their patients' DRBW ratios.

3.2 The Clinical Standard for Allograft Size Matching and Concerns for Mismatch

The current standard of care for HTx centers to determine if an offered allograft will appropriately fit a patient is to use the DRBW ratio as a standalone metric [46]. The offered donor's weight is the value placed in the numerator position and the recipient's weight is placed in the denominator position. In general, a DRBW ratio matching 1 is considered preferable [57–59].

The lower and upper weight range centers are potentially willing to accept for their patient, is required to list their patient onto the UNOS database. The listing standard of care is for clinicians to use the DRBW ratio methodology as a tool to identify the weight listing range they are potentially willing to accept. To ensure a patient is provided with allograft offers that are size appropriate while not unnecessarily limiting their donor pool, centers will set a lower and upper DRBW ratio criteria range they are willing to consider. Due to the donor shortage, many centers are progressive in allograft procurement for their patients by widening the DRBW range criteria they list their patients for and then, within reason, are willing to accept size mismatched donor offers [59–64]. Pediatric

centers are generally more aggressive with the lower and upper DRBW ratio ranges they are willing to set due appropriate size allografts being grossly limited for their patient population [59,61]. Typically, centers will only accept offers that approach the outer boundaries of the listed DRBW ratio criteria when there is concern that another, more appropriate donor offer will not be made available before the patient reaches a level of deterioration that is irreversible.

The DRBW ratio criteria that patients are listed for are center specific and can occasionally be age-specific. Conventional centers will, in general, set the criterion range in pediatric patients from 0.75 to 1.50 and increase the criterion range from 0.75 to 3.00 for infants that are ≤18 months old [46]. A key clinically contributing factor to the increased DRBW ratio range in infants is the donor allograft shortage in which approximately 25% of listed infants will die on the waitlist [6,59,61]. A select group of centers (including PCH) have historically expanded the DRBW ratio range from 0.70 to 4.00 in extreme cases. Although clinicians expand the DRBW ratio range to help ensure they procure an allograft for their patient, clinical complications arising from cardiac allograft size mismatches are a concern.

First, undersized allografts affect the cardiac output the allograft can perform (i.e., $Cardiac\ Output = Stroke\ Volume * Heart\ Rate$). Cardiac output (flow rate) is the systemic volume of blood the heart can move over a given period of time. Flow rate needs are proportional to an individual's body size, i.e., adults have a larger cardiac output need then infants [37]. A heart being able to meet an individual's cardiac output needs, relative to size, is an important requirement of a well-functioning organ, as is demonstrated with hemodynamic monitoring systems reporting a patient's cardiac index [65].

An undersized allograft would need to increase its heart rate to provide an appropriate circulatory flow rate. *Hosenpud et al.* published empirical results supporting increased heart rate is a likely consequence of allograft undersized mismatch [57]. This increased heart rate would add stress to the allograft that could potentially have negative clinical consequences [66]. An excessively undersized allograft could potentially be incapable to meet a patient's cardiac output needs. In fact, clinical practice is to avoid undersized allografts (preferring slightly oversized allografts) for patients with elevated pulmonary vascular resistance due to concern for

hemodynamic insufficiency and the potential development of right ventricular failure [58,63,66]. The concern for right ventricular failure development in the undersized allograft is a potential consequence of the long-term lung hemodynamic resistant load the inferior muscle must contend with to drive the pulmonary circulation.

Second, oversized allografts are likely to compress the surrounding anatomical structures and themselves [59,67]. Furthermore, there is concern that an oversized allograft will over-perfuse the patient [59,67]. The various types of compression effects clinicians are weary of in accepting an oversized allograft include lung compression, lung collapse, descending aorta compression, and pulmonary vein compression [59,61,67]. An allograft "self-compressing" itself is one concern that can compromise circulatory function [67]. To avoid "self-compression" the surgeon will slowly reduce the sternum chest spreader while intensely monitoring the pressure of the right atrium in the operating room. In situations of sudden, excess right atrial pressure (or other clinical indications of oversized compression effects) during attempted chest closure, the chest will remain open and closure will be completed at a later date – opening one or both pleural cavities have also been performed when needed [59,67]. Sudden over-perfusion is a potential complication of transplanting an oversized allograft in which a sudden increase of blood flow to the brain can cause neurological symptoms. The neurological symptoms of oversized allograft mismatch include coma, convulsions, and headaches but these issues normally resolve after several day [59,67]. Although oversized allografts might be well-tolerated, as is suggested with the expanded DRBW ratios HTx centers will accepted, excessively oversized allografts are a clinical concern for HTx recipients.

3.3 Limitations of the DRBW Ratio Method and other Previously Proposed Solutions

The DRBW ratio metric used in allograft size matching in HTx assumes that, in general, similar body weights correspond to similar patient and donor TCVs and that the offered allograft is capable of sustaining appropriate hemodynamics long-term [57,58,66,68]. Although the DRBW ratio metric is the current standard of care, there are inconstancies on whether this ratio predicts patient fit-

related outcomes well [69–72] or poorly [53,55,56,59,61,62,73,74], or fit-related outcomes depends on patient size [66,75].

The conventional pediatric DRBW ratio criteria that centers typically list for may be derived from adult HTx studies and may explain why pediatric studies are suggesting that DRBW ratio ranges can be expanded [46,49,54,56,61,67,74,76]. Referring back to chapter 2, it was postulated that TCV and body have a linear, isometric scaling relationship (i.e., $\hat{b} \approx 1$) in the adult population. This postulated adult scaling relationship was found by the author – to be presented in chapter 4 – to not hold when pediatric developmental growth was involved (i.e., $\hat{b} \approx 0.84$). As discussed in chapter 2, there are ample examples for empirical scaling exponent relationships found in adult mammals to change when developmental growth is involved – including between adult and pediatric developmental growth in human cardiac anatomy and physiology [7,21,23,24,27,37,40,77–80]. Applying adult-derived DRBW ratio boundary limits to pediatrics imposes the adult scaling relationship (likely an isometric scaling relationship) onto pediatric patients during allograft size-matching.

The decrease in the scaling exponent between the adult postulated value and the measured value of the pediatric population mathematically suggest adult DRBW ratio criteria are limiting pediatric donor pools. The mathematical argument that adult criteria are limiting the donor pool for pediatric patients is based on the clinical assumption that the upper and lower donor-recipient TCV ratio are needed to be held constant between these age groups and not the DRBW ratio for a safe allograft fit. The smaller scaling exponent in pediatrics mathematically implies a larger weight range is needed so the TCV ratio matches the adult population.

To demonstrate this statement, let us consider the historical DRBW cut-off range of $\pm 20\%$ that was often suggested (corresponds to a DRBW ratio range of 0.80 to 1.20) [55,59]. Based on (1) the postulated adult scaling exponents and (2) the mathematical properties of exponents, the corresponding lower and upper TCV ratio ranges for the adult can be calculated from the historic DRBW ratios as follows:

$$Lower\ TCV\ Ratio = Lower\ DRBW\ Ratio^{1} = 0.80^{1} = 0.80$$

and

$$Upper\ TCV\ Ratio = Upper\ DRBW\ Ratio^{1} = 1.20^{1} = 1.20$$

Now, using our previously listed assumption that the lower and upper TCV ratios safe ranges are the same constants for both adult and pediatric populations, let us find the corresponding lower and upper DRBW ratio criteria for pediatrics (using the pediatric scaling exponent):

$$Lower\ TCV\ Ratio = 0.80 = Lower\ DRBW\ Ratio^{0.84}$$

$$\rightarrow$$

$$Lower\ DRBW\ Ratio^{0.84/0.84} = Lower\ TCV\ Ratio^{1/0.84} = 0.80^{1/0.84} = 0.77$$

and

$$Upper\ TCV\ Ratio = 1.20 = Upper\ DRBW\ Ratio^{0.84}$$

$$\rightarrow$$

$$Upper\ DRBW\ Ratio^{0.84/0.84} = Upper\ TCV\ Ratio^{1/0.84} = 1.20^{1/0.84} = 1.24$$

The mathematical results suggest a wider DRBW ratio range is possible in pediatrics if matching the TCV ratio range is the clinically relevant factor. Although this increased weight range is minimal it becomes more pertinent if one is trying to match a TCV ratio of 3.00 - this corresponds to a DRBW ratio of 3.57 in pediatrics. Applying adult DRBW ratio criteria fails to consider developmental growth and this mathematical exercise suggest adult criteria may potentially be limiting pediatric donor pools unnecessarily.

Abnormal body weight (e.g., obesity and excess fluid overload) may further negate the clinical relevance of the DRBW ratio metric [46,66,67,81,82]. Excess fluid overload is common in waitlisted patients as a complication of end-stage congestive heart failure [46,66]. To overcome abnormal body weight limitations, recent publications have suggested potential applications of echocardiography-derived estimates to predict cardiac fit [46,81,83]. Some centers have suggested height is a more appropriate then weight is in the DRBW ratio metric because the height measurement is not generally susceptible to edema or subcutaneous fat [46,81,83,84].

Pathology and progression of disease further complicate allograft size-matching based on the DRBW ratio as a standalone metric. For example, size-matching allografts for patients with congenital heart disease is nontrivial [67,85]. Some clinicians and clinical research groups suggest patients with enlarged hearts, e.g., dilated cardiomyopathy, can tolerate an allograft with a DRBW >> 1, i.e., well-oversized allograft [46,54,59,67,81,83]. In fact, PCH will take linear measurements of chest x-ray and compute the cardiothoracic ratio for patients with cardiomyopathy to potentially further expand their patients' DRBW listing ranges, and thus further expand their patients' listed donor pools.

3.4 Proposal of a Virtual Heart Transplant Fit Assessment Tool

The previous chapter section discussed that although the DRBW ratio is the clinical standard to predict if an allograft will fit within a patient, it is an indirect size matching metric that is far removed from the cardiac anatomy and susceptible to abnormal weight. Furthermore, it is standard for clinicians to expand the lower and upper DRBW ratio criteria they are willing to accept to ensure allograft procurement for their patient. However, by expanding the DRBW ratio, clinicians are reducing the safety margins of this metric when there are already established limitations with this method.

To address concerns of expanding donor pools by expanding DRBW ratio criteria, some groups have looked for solutions outside of weight comparisons. Recently proposed solutions to assess donor-recipient allograft match were presented in the previous chapter section but these

solutions have their own set of limitations. Height has been suggested to address the abnormal weight limitation but it is still far removed from the cardiac anatomy. While echocardiography-derived and x-ray-derived fit assessment methods are more direct, they may be limited in accounting for both donor and/or patient cardiac and thoracic cavity aberrations.

Considerations of a congenital heart disease patient's unique three-dimensional (3D) cardiothoracic space will need to be taken into account to safely expand a patient's donor pool. Congenital heart disease patients are cases with surgically complex thoracic environments that account for approximately 50% of infant and child HTx cases and 25% of adolescent HTx cases [72,85–90]. If a novel tool is to be successfully developed to safely expanding donor pools based on size matching in the pediatric HTx center then it will need to consider patient-specific anatomical aberrations of the cardiothoracic-space.

Fitting ventricular assist devices and artificial hearts based on patient size is an area of medicine analogous to cardiac allograft sizing matching. There has been a recent push to safely assess mechanical device fit by performing computer fit assessments, i.e., virtual implant assessments [91–93]. A virtual implant assessment is performed by strategically fusing a medical device's geometry onto a patient's medical image. These case studies have shown instances where large mechanical circulatory support devices fit in undersized patients failing to meet standard fit criteria assessments [91–93]. The potential benefit of a virtual implant assessment is a patient's structural aberrations can be accounted for when fusing a computer geometry of an implant device onto the patient's medical image. These case studies indicate there are pediatric subpopulations that may in fact have the thoracic morphology capable of supporting a larger donor heart. The work herein leverages the virtual implant assessment techniques for allograft size fit assessment.

This study hypothesizes that current computer technology can be leveraged to better assess a donor's fit within a patient's anatomy by computationally simulating a transplant surgery in the computer, i.e., a virtual transplant. It is speculated that virtually fitting, i.e., fusing, an allograft onto a pre-operative recipient medical image dataset to assess for potential fit-related

complications will, in general, address the limitations of the DRBW and echocardiography-derived methods.

The virtual assessment will involve taking either a donor's cardiac reconstruction or a cardiac reconstruction from a healthy patient with a matching TCV. These reconstructions are of the TCV in which blood volume and tissue that is deep to the myocardial exterior surface are included. A surgeon will then be able to mimic a surgery by translating and rotating the allograft geometry into place – hence the term "virtual surgery". The surgeon will be able to move the allograft until they determine either (1) a fit is safely possible or (2) a fit is not safely possible.

Although it might be considered preferable to use a donor's allograft reconstruction for virtual HTx fit assessment, a regression model to predict allograft TCV and a library of healthy heart reconstructions will need to be developed for virtual fit assessments to be clinically feasible. A regression model to predict allograft TCV from gross donor metrics and library of healthy heart geometries will be needed for the following reasons: (1) donors will rarely have CT or MR medical images available and (2) time constraints, e.g., image transfer, reconstruction, and virtual fit assessment times. It is therefore hypothesized that a regression model to predict allograft TCVs can be developed from non-invasive donor parameters readily available to a recipient's HTx center and validated.

Once the novel tool is developed, driven by a developed regression model, a fit assessment could potentially be achieved within the allotted time a center usually has for deciding to accept an allograft offer or not. In real-time the clinical team will be able to virtually assess an offered allograft's fit in a patient's complex, congenital thoracic environment. A clinician would be able to visualize the fit and decide to accept an allograft that may have traditionally been considered too large. This allows the clinician to expand their patient's donor pool in real-time. Finally, it is hypothesized virtual fit assessments will be demonstrated to help safely expand the patient's donor pool.

3.5 Types of Heart Transplants

Chapter 3 will now be concluded with a technical overview of the different types of HTx procedures that a surgeon can perform. The purpose of the overview is to discuss how the technical differences of each surgical procedure might affect how the final virtual assessment is developed and/or implemented. There are 4 distinct HTx procedures clinically available. These distinct procedures are classified into one of two transplant types: (1) heterotopic heart transplant (HHT) and (2) orthotopic heart transplant (OHT). The OHT class encompass 3 of the 4 district types of HTx in which the diseased, native heart is removed. The HHT class is unique because the diseased, native heart remains and therefore a patient will have two functional hearts - although the native heart has limited functionality. All 4 types of HTx are unique at the procedural level because of the variations in how the allograft can be anastomosed to the patient.

As a reminder, a virtual fit assessment is performed by taking a healthy allograft geometry and fusing it onto a patient's medical image for assessment. The assessment espies any overlap of the allograft with surrounding anatomical structures to suggest if there are clinical concerns for compression effects.

HHT is largely an outdated technique in which the donor heart is attached to the failing native heart [94–96]. In general, only when a patient is offered an excessively undersized allograft and/or the patient has high pulmonary resistance a clinical team in the USA might consider performing a HHT in the current medical era [94–96]. The typical procedure is to anastomose the right atriums and the ascending aortas of the donor and recipient together [94–96]. The result is a HHT recipient has two functioning hearts – although one is a failing heart. There are cases in which HHT can be used to let the disease heart recover but mechanical support appears to be preferable for such circumstances in the current era [95–98]. The current virtual assessment tool has potential to be used to assess HHT fits by assessing how the donor heart would fit within the patient's right plural cavity. The surgeon performing the virtual fit would need to carefully rotate the allograft geometry so the right atriums are near one another. Approximate placement of the allograft's right atrium to the recipient's right atrium is required for the structures to be anastomosed. However,

there is yet to be a HHT procedure performed at PCH and therefore it will not be investigated herein.

An important observation of HHT procedures is the clinical feasibility to fit a second, although smaller, heart into a patient's chest cavity. This largely outdated clinical practice therefore suggest there are circumstances in which an oversized allograft can safely fit a patient. Since the HHT procedure can fit a secondary heart into a patient's cavity, it is further justification that a virtual assessment tool can be developed to assess oversized allograft fit – as developed herein.

OHT is now the gold standard in end-stage heart failure treatment after the development of immune suppressant mediations [94–96]. Suppressants help to prevent or at least minimize allograft rejection Before suppressants were available HHT was preferred because at least the native heart could potentially be preserved if the donor allograft was rejected [95,96]. In general, OHTs are a class of procedures in which the vast majority of the patient's heart is removed. Only a small patch or set of patches of the native heart are kept to anastomose the donor heart to the patient [94,99]. The invasiveness of OHT transplants make them a permeant surgical modification. The current OHT transplants are broken up into 3 types: total, bicaval, and biatiral transplants [94,99,100].

First, total OHT is a procedure in which 6 anastomoses are to be made at the superior vena cava, inferior vena cava, pulmonary artery, aorta, and at 2 patches encompassing either the left or right pulmonary veins' inlets to the heart [94,99]. This procedure preserves both atrium and thus helps to protect both the mechanical and electrophysiology of the donor heart [99–102]. Although this procedure has preferred outcomes (when done correctly), it is technically challenging with a particular concern of correctly attaching the pulmonary veins to the donor heart without bleeding [94,99,102,103]. Attaching already small vessels, e.g., pulmonary veins, can potentially further complicate the procedure in the smallest of children. The technical challenges of pulmonary vein attachment make the remaining 2 OHT methods preferable because they are associated with much more clinically manageable complications.

Second, bicaval OHT is a procedure in which 5 anastomoses are made [94,99]. The anastomosis of the bicaval procedure are similar to the total OHT except there are no pulmonary

vein anastomosis. Instead of attaching the pulmonary veins to the donor allograft a single patch comprised of the back of the recipient's left atrium (which contains the pulmonary vein) is anastomosed to the allograft. Bicaval OHT is clinically preferable as it is a trade-off between minimizing post-operative mechanical and electrophysiology complications by preserving the right atrium and the relative ease for the surgeon to safely implement the anastomoses [94,99,100,104].

Finally, biatrial OHT is a procedure in which 3 anastomoses are to be made at the pulmonary artery, aorta, and a single patch of the posterior atrium wall (containing both the left and right chambers) that encompasses the vena cava and pulmonary veins [94,99]. This procedure is associated with post-operative complications that include mechanical, electrophysiology, and thrombosis issues [99,104]. However, anastomosing 3 large structures is technically simplistic, making it the easiest OHT to surgically implement, in general. In fact, even though biatrial OHT is associated with higher risk for post-operative complications, it is the preferred transplant in (1) infants and the smallest of children and (2) for highly complex structural congenital heart disease cases [90]. The few, large anastomoses sites greatly ease the surgeon's ability to safely suture the allograft to the patient.

In all three OHT procedures the donor heart is manipulated such that the allograft can be anastomosed to the patient. The total OHT leaves the donor heart with minimal manipulation and therefore more effectively parallels the virtual fit assessment procedure because the allograft TCV (with no geometric manipulation) is fused to the recipient medical image. While bicaval and biatrial techniques involve relatively more complex modification to the donor allograft then a total OHT, all three procedures leave only small patches of the patient's native heart to anastomose to the allograft.

To gauge if there was enough clinically perceived value in modeling the virtual allograft geometries for all 3 types of OHT, the author talked to cardiologists and cardiovascular surgeons at PCH. Although there was high interest in developing a tool to assess HTx fits with the allograft intact, there was no interest in manipulating the allograft geometry for any of the 3 types of OHT. In fact, there a very strong and resounding negative response from the clinical community to investigate fits in which slight modifications were made to the allograft geometry. Dr. Tara

Karamlou, a cardiothoracic surgeon, summed up the reason for the clinical disinterest in assessing virtual fits based on specific OHT procedures as follows, "[…] the anastomoses for OHT are non-space-occupying lesions". She continued to refer to the anastomosis cuffs cited in literature as "patches" to further convey the anastomosis sites had little to no effect on final transplanted heart volume. For the author, Dr. Karamlou's use of the word "patch" was a powerfully descriptive explanation that the anastomose cuff had perceived negligible effect on the final transplanted volume and therefore the author has retained the term herein.

As demonstrated by Dr. Karamlou's feedback on how to perform the virtual HTx fit assessments, the clinical perception is the anastomoses required in OHT result in a zero-net TCV gain/loss for the donor allograft. Given the current clinical disinterest in assessing donor volumetric changes due to procedural anastomoses, this area of consideration was not considered as part of the work herein. If future virtual HTx fit assessments start to either consider (1) flow, thrombus, and/or compression complication effects through fluid-surface interaction simulations or (2) if it becomes possible to simulate electrophysiology changes as the allograft adapts and ages post-HTx, then simulating procedures based on OHT type should be investigated.

CHAPTER 4

DEVELOPMENT OF THE HEALTHY HEART LIBRARY

The design concept of the novel virtual HTx fit assessment tool required a dataset of normal, healthy heart computer geometries, i.e., reconstructions, with the associated noninvasive, gross parameters included. The dataset was referred to as the “healthy heart library” or informally as the “library”. The library was used for 2 distinct tasks in developing and analyzing the novel tool. First, the library was used to develop a regression model that predicts a donor’s allograft TCV from gross donor parameters, i.e., predictors. Second, the TCV reconstructions were used to assess if an offered allograft would fit by performing a geometry-driven virtual fit assessment. The geometry-driven assessment was achieved by using a library reconstruction that matched the donor’s predicted allograft TCV.

Chapter 4 covers the development and the demographics of the healthy heart library. The library development, e.g., data collection, was performed via a retrospective, chart review. The chart review isolated patients with normal, healthy cardiac anatomies and physiologies that were available in CT or MR images. TCV reconstructions were generated from the acquired medical images and then had their volumes measured. CT and MR measured TCV (mTCV) values were statistically analyzed to help ensure minimized or eliminated bias between the modality measurements. The lack of a mTCV modality bias would justify combining both the CT and the MR data into the allograft TCV prediction model and, in general, the tool’s overall development process. The allometric scaling relationships between TCV and gross geometric predictors within the library were analyzed to appreciate both the data and the modeling process undertaken in chapter 5. For this body of work, the creation of the healthy heart library dataset constituted as Aim 1.

4.1 Methods and Materials of the Healthy Heart Library

The IRB approved a chart review for 100 (minimum 50) normal, healthy heart subjects with retrospective CT or MR images. Pulled subject data was de-identified. The subject data was used to develop the allograft TCV prediction model and the TCV reconstructions.

The requested subject sample size range was selected to meet the typical statistician recommendations for developing a parametric, linear regression model. The typical recommendations for parametric, linear regression modeling are that (1) there is at least 10-20 data points per potential predictor parameter and (2) there is no less than 50 data points in total [105]. From a practical point-of-view, the sample size cutoff was kept to 100 subjects because there was a concern there would be a limited number of healthy heart patients available in the hospital's chart system.

The healthy cardiac subject parameters were either pulled during the chart review, calculated from the pulled parameters, or reconstructed and measured from the pulled medical images. The data and their units (or classification levels) that were included into the healthy heart library were listed in Table 4.1. Body surface area (BSA) was calculated using the Mosteller equation [106]. Body mass index (BMI) was calculated using the specific relative body weight indexing equation that was originally coined to that same clinical term, i.e., "BMI" [107]. The mTCV and the gross parameters (i.e., Sex Age, Ht, Wt, BSA, and BMI) were specifically included within the library to serve as the model's criterion variable and as potential model predictors, respectively.

Healthy Heart Library Data

Chart Data:	***Classification or Units***
Cardiac Medical Images	CT/MR
Sex	Male/Female
Age	months
Height	cm
Weight	kg
Calculated:	
Body Surface Area	m^2
Body Mass Index	kg/m^2
Reconstruction:	
Total Cardiac Volume Geometry	-
Measured:	
Total Cardiac Volume	mL

Table 4.1: The list of chart data pulled from healthy heart subjects were presented in the current Figure. BSA and BMI were calculated from Ht and Wt. TCV was reconstructed and measured from cardiac medical images.

Identification of normal, healthy heart subjects were performed by reviewing radiologist reports. Subjects reported as having no greater than mild dilation or obstruction of the greater vessels or coronaries were considered to be included in the library. Greater vessel and coronary anomalies deemed to not affect either the cardiac anatomy (including size) or physiology were included within the library as normal. Subjects with pathologies or have undergone treatments known to potentially affect cardiac anatomy, size, and pathology were automatically excluded. A cardiologist involved in the study confirmed the normal, healthy heart findings before including the subject into the dataset. Subjects that were diagnosed with anemia or had undergone some classes of chemotherapy drugs were typical examples where the cardiologist had automatically excluded the individual from the health heart library.

Reconstructions of subject healthy hearts were generated from CT or MR cardiac image volumes. When possible, the image volumes at cardiac peak diastole were used because the ventricular muscles would be relaxed, i.e., the ventricles would be at or near their volumetric largest. Only the TCVs were reconstructed for this body of work. The reconstructions were produced in the following sequential order: (1) images were segmented in Mimics Innovation Suite software (Materialise, Leuven, Belgium), (2) initial reconstructions were post-processed in Geomagic Studio software (3D Systems, South Carolina), and (3) the completed reconstructions were fused to the subject's images within Mimics for quality assurance. The myocardium and anatomical structures deep to the myocardium exterior surface (including blood volumes) were defined as part of the TCV. Figure 4.1 shows a reconstruction of a subject's TCV with the reconstruction's contour fused onto the medical image. The mTCV values were acquired from the reconstructed geometries using Mimics.

Healthy Heart Reconstruction Showing TCV Boundary

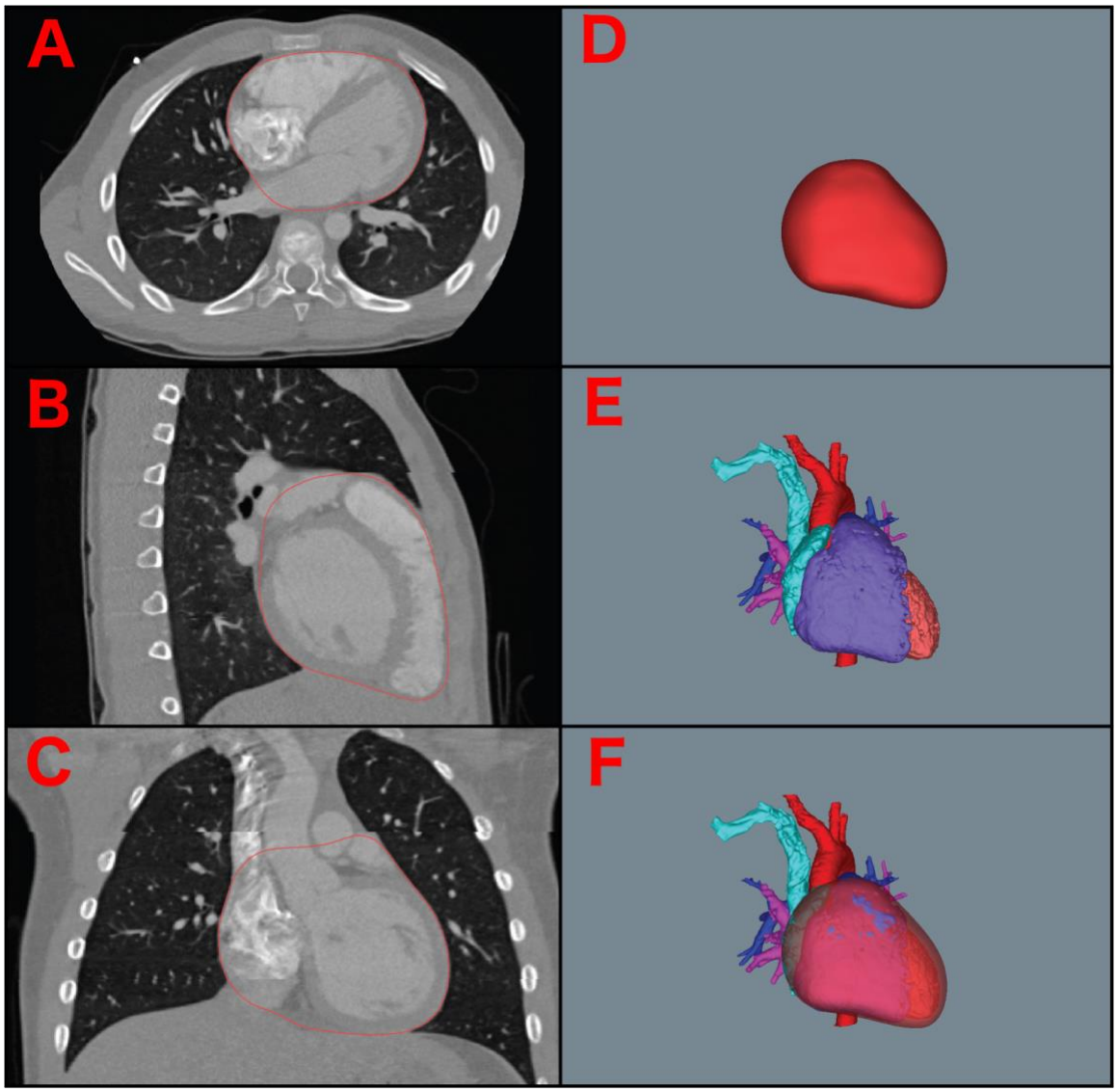

Figure 4.1: Visualization of a healthy TCV reconstruction (red). TCV contour lines were fused onto the subject's orthogonal medical image viewing planes (A, B, and C) to illustrate the TCV boarder – the reconstruction included blood volumes (image (E)) and anatomical structures deep to the myocardium exterior surface. The 3D TCV geometry were illustrated with (F) and without (D) reconstructed blood volumes independently reconstructed. The transparent setting of the TCV in image (F) demonstrated the thick and thin exterior muscular walls of the left and right ventricles.

The mTCV values were statistically tested to ensure there was no significant difference between the CT and MR measurements. Failing to detect a statistical difference was perceived to suggest the modality measurements were equivalent and provide justification for combining the mTCVs within the healthy heart library. Suggesting that the mTCVs were equivalent would support

the allograft prediction model, the recipient images, and donor allograft geometry do not need to account for modality. The statistical test was performed with a one-way ANOVA (alpha = 0.05, $H_0 = 0$) in the commercial software JMP (SAS Institute Inc., Cary, North Carolina); the null hypothesis was the modalities had equivalent mTCVs.

The allometric scaling signals for the library's geometric predictors were analyzed to appreciate the upcoming TCV modeling process and more generally the TCV signals related to developmental growth. Specifically, the geometric signals analyzed were between TCV and the geometric-specific gross donor parameters (Ht, Wt, BSA, and BMI). The allometric signal analysis consisted of first natural log transforming both the TCV and gross parameters and then performing a series of 4 univariable, linear regression fits. The allometric signal fitting process was repeated twice in which (1) all 97 data points were used and (2) Cook's distance ($4/N$; were "N" was library size) removed influential outliers [108,109].

4.2 Results of the Healthy Heart Library

A total of 97 healthy heart subjects were identified and included in the healthy heart library dataset. There was a bias towards the male sex (64%) in the library dataset. Library demographics were presented in Table 4.2. Figures 4.2 illustrated the healthy heart library subjects' mTCVs verses their corresponding Sex, Age, Ht, Wt, BSA, and BMI parameters. Figure 4.3 was a BMI color-coded version of Figure 4.2 and included to illustrate the interaction effect between BMI, i.e., body type, and the corresponding predictors on TCV.

Demographics of the Healthy Heart Library (N=97)

Demographic characteristics	Results
Heart Volume (mL)	535 ± 259 (36 - 1340)
Male	62
Female	35
CT	50
MR	47
Age (months)	155 ± 70 (3 - 358)
Height (cm)	147 ± 31 (42 - 186)
Weight (kg)	53 ± 29 (2.8 - 139)
BSA (m^2)	1.46 ± 0.54 (0.18 - 2.59)
BMI (kg/m^2)	22.8 ± 7.2 (13.5 - 46.5)

Data are reported as mean ± standard deviation (range) or N

Table 4.2: Normal cardiac subject demographics used to train and test the predictive models in chapter 5. Although imaging modality was not a predictor in the regression model, the mTCVs were derived from the CT and MR images and therefore the modality demographics were included. Figure 4.4 and corresponding ANOVA test (presented shortly) showed CT and MR measurements were equivalent and therefore justified the combining of both modalities' mTCVs and reconstructions for use in the virtual HTx fit assessment tool.

Plot of Healthy Heart Library Demographics (N=97)

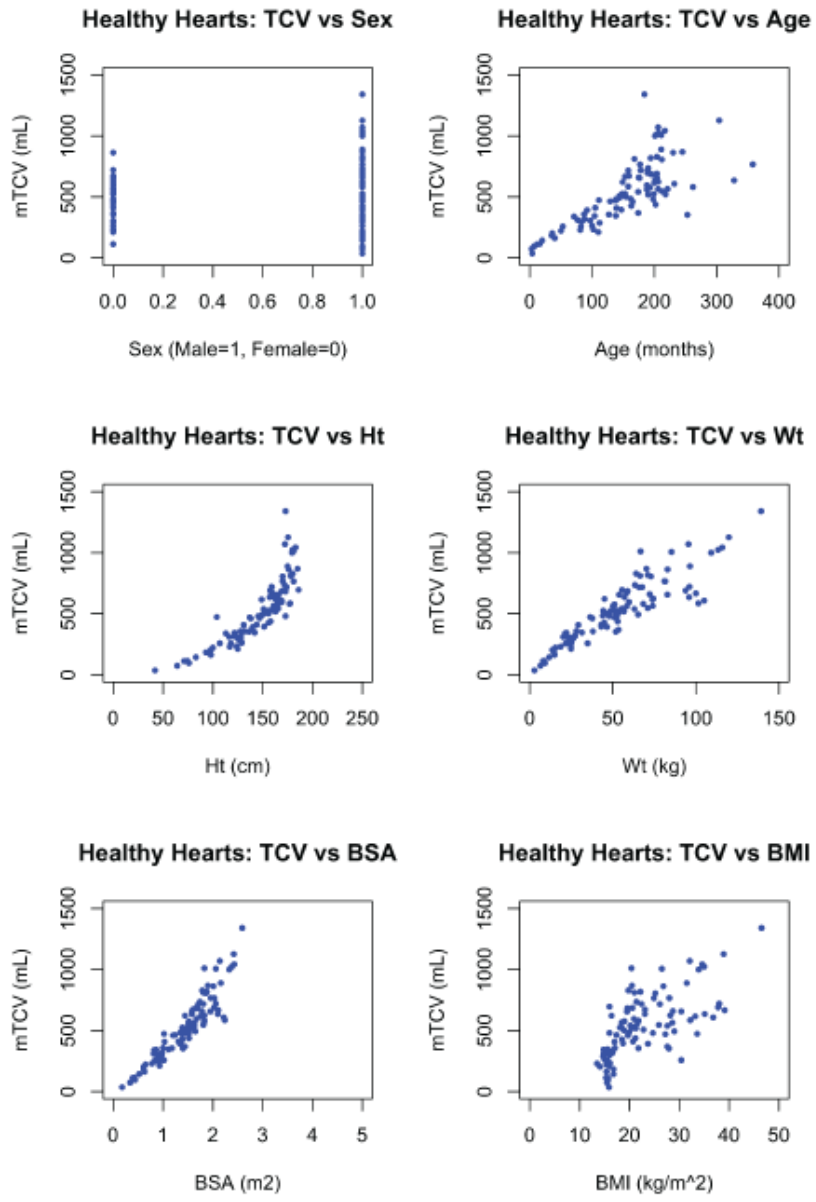

Figure 4.2: Healthy heart parameters plotted against mTCV for all 97 subjects. Visual inspection suggested nonlinear and possibly linear trends were present in the data between subject gross parameters and TCV. The nonlinear trends suggested the data will need to be transformed for linear regression techniques to be implemented in chapter 5's linear TCV modeling process. The data trends suggested growth patterns can be leveraged in an allograft TCV prediction model.

Color-Coded Plot of Healthy Heart Library Demographics (N=97)

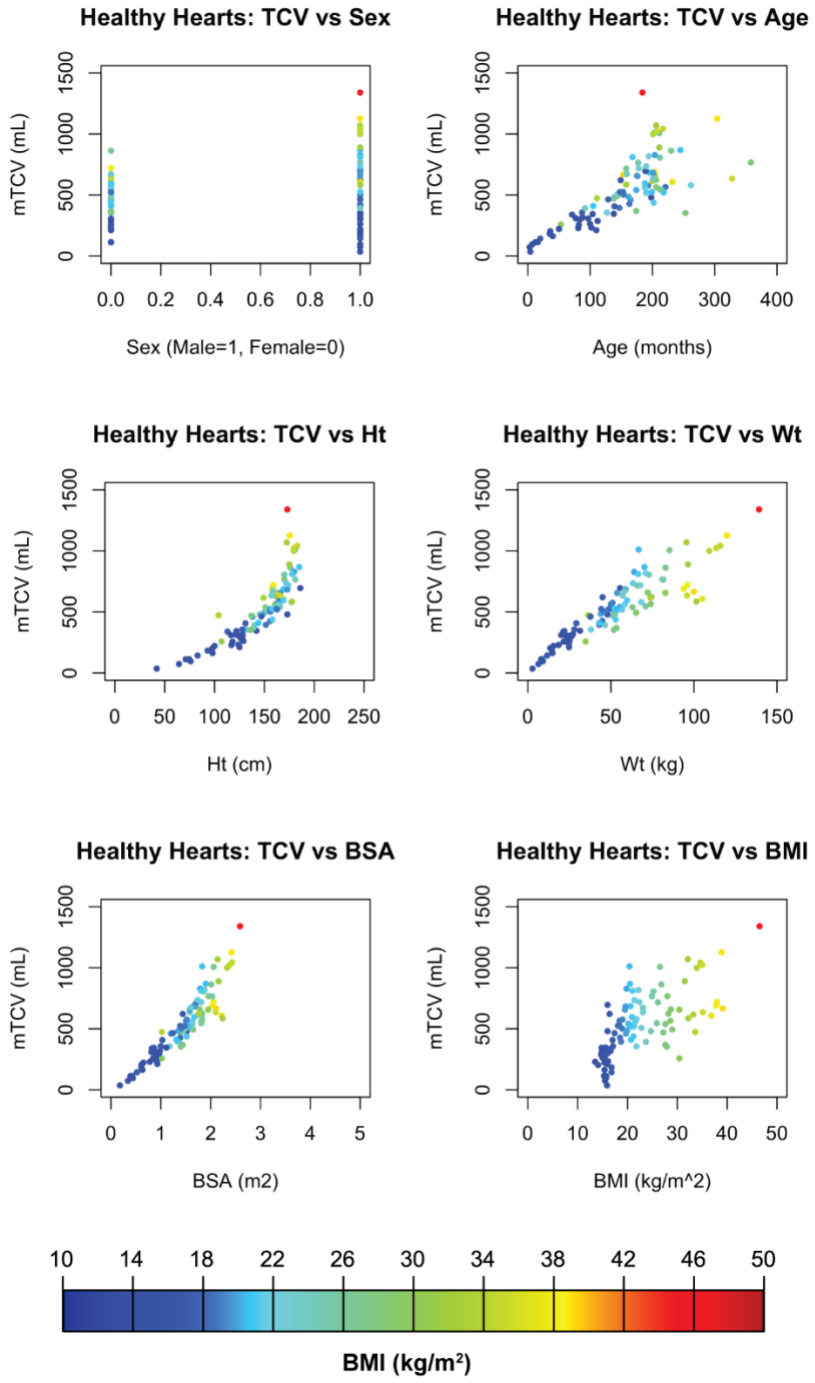

Figure 4.3: A BMI color-coded plot of Figure 4.2 suggested body type, i.e., obesity, along with developmental growth, can be leveraged in TCV modeling process. Visual inspection suggested there were influential outliers in the data, e.g., a morbidly obese individual (marked red).

Subject demographic percentages were grouped by age: 6% (infants and toddlers, 0-2yrs), 29% (children, 2-12), 60% (adolescents, 12-21), and 5% (adult, 21+) [110]. Breaking the age group into groups by year for 0 to 21 years old, i.e., 22 groups, resulted in each of the age groups representing, in general, between 1% and 5% of the library's total data population. The adolescent age groups spanning 12 to 17 were the exceptions in which each of the age groups represented between 6% and 11% of the total data population. 95% of the subjects were of the adolescent age or younger at the time of their image scan. The oldest individual in the library was 29 years old at the time of their image scan.

Subject demographic percentages grouped by body type, i.e., BMI, were: 43% ($BMI < 20\,kg/m^3$), 25% ($20\,kg/m^3 \leq BMI < 25\,kg/m^3$), 14% ($25\,kg/m^3 \leq BMI < 30\,kg/m^3$), and 18% ($30\,kg/m^3 \leq BMI$). BMI $mean \pm St.Dev.$ by age group were: $16 \pm 1\ kg/m^3$ (infants and toddlers, 0-2yrs), $18 \pm 5\ kg/m^3$ (children, 2-12), $25 \pm 7\ kg/m^3$ (adolescents, 12-21), and $30 \pm 7\ kg/m^3$ (adult, 21+) [110]. Figure 4.4 illustrated the relationship between Age and BMI.

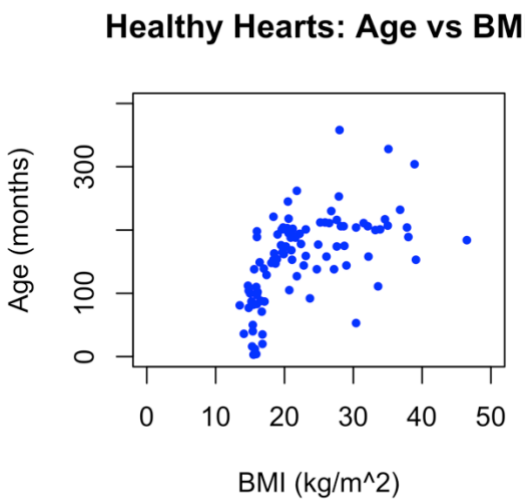

Figure 4.4: Visual graph inspection suggested subject BMI was generally independent of subject Age. The exception appeared to be for subjects < 100 months of Age in which their BMI was held relatively constant.

Image modality demographics had a near negligible bias towards the CT modality (52%) in the healthy heart library. The ANOVA test failed to reject the null hypothesis that the mTCVs

were statistically the same between image modalities (p-value = 0.5156, Cohen's distance effect size = 0.1328). This ANOVA result suggested the CT and MR mTCVs were equivalent in the practical sense. Figure 4.5 visually illustrated the ANOVA results in a box-and-whisker plot.

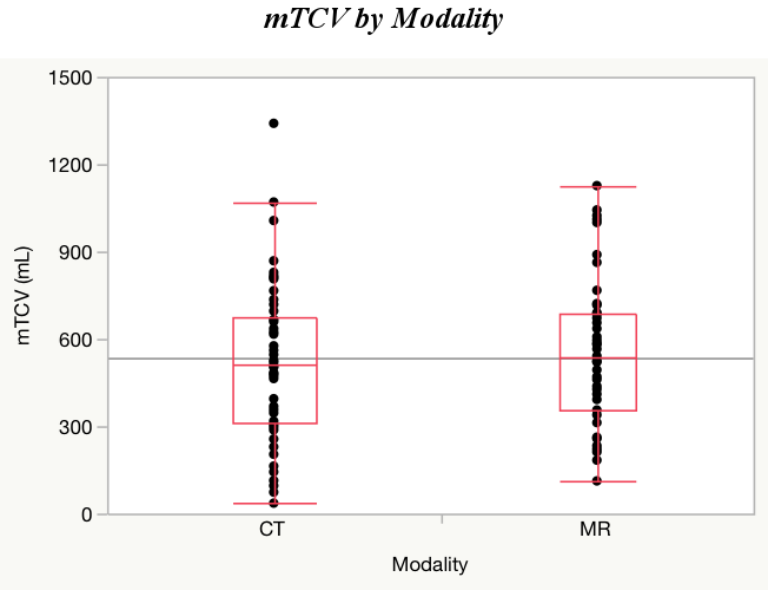

Figure 4.5: mTCVs for the different modalities. The small box-and-whisker plot horizontal lines (red) indicate the lower whisker, 25th-quartile, 50th-quartile, 75th-quartile, and upper whisker for CT and MR. The 25th-quartile, 50th-quartile, 75th-quartile values were (314mL, 513mL, and 674mL) for CT and (356mL, 539mL, and 685mL) for MR, respectively. The large horizontal line (grey) indicated the grand mean of 535ml, i.e., the mean for the N = 97 samples.

Allometric scaling signals between geometric-specific parameters and TCV were estimated and presented in Tables 4.3 and 4.4. The estimated coefficients and standard errors (SEs) were determined with (Table 4.4) and without (Table 4.3) influential outliers removed. In general, minimal changes in estimated coefficients and SEs were observed, with BMI being the exception. The scaling exponents were compiled in Table 4.5 and generally showed the estimated, i.e., empirical, values were slightly less than the postulated, i.e., theoretical, values.

The Pearson's correlations for the TCV and the continuous predictors in the library were presented in Table 4.6. The correlation between TCV and BMI was a particularly weak (0.67). The correlation between TCV and Ht, Wt, and BSA were stronger (> 0.84).

Allometric Scaling Signals of Geometric Metrics with TCV: With Influential Outliers

Structural Framework (R software notation)				
TCV ~ 1 + Variable				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \mathfrak{a} + \mathfrak{b} * \ln(\text{Variable})$ ***or*** $\text{TCV} = \mathbf{e}^{\mathfrak{a}} * \text{Variable}^{\mathfrak{b}}$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\mathfrak{a}_{\mathbf{Ht}}$	-5.07466	0.3922915	-12.93594	0
$\mathfrak{b}_{\mathbf{Ht}}$	2.258016	0.0789381	28.60488	0
$\mathfrak{a}_{\mathbf{Wt}}$	2.9793897	0.10269996	29.01062	0
$\mathfrak{b}_{\mathbf{Wt}}$	0.8306572	0.02660205	31.22531	0
$\mathfrak{a}_{\mathbf{BSA}}$	5.777092	0.01933058	298.85762	0
$\mathfrak{b}_{\mathbf{BSA}}$	1.246316	0.03473741	35.87819	0
$\mathfrak{a}_{\mathbf{BMI}}$	1.993526	0.5139927	3.878511	2.00E-04
$\mathfrak{b}_{\mathbf{BMI}}$	1.343744	0.1661563	8.087231	0.00E+00

Table 4.3: The 4 allometric scaling coefficients, i.e., $\mathfrak{b}$ values, between TCV and Ht, Wt, BSA, and BMI were calculated from the library (N = 97). As a mathematical consequence of how the univariable regression was fitted (the input variables were natural log transformed before fitting) the $\boldsymbol{a}$ coefficient in the generic allometric equation was $\boldsymbol{a} \triangleq \mathbf{e}^{\mathfrak{a}}$, i.e., $\mathfrak{a} = \boldsymbol{ln}(\boldsymbol{a}) \neq \boldsymbol{a}$.

Allometric Scaling Signals of Geometric Metrics with TCV: Without Influential Outliers

Structural Framework (R software notation)				
TCV ~ 1 + Variable				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \mathfrak{a} + \mathfrak{b} * \ln(\text{Variable})$ ***or*** $\text{TCV} = \mathbf{e}^{\mathfrak{a}} * \text{Variable}^{\mathfrak{b}}$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\mathfrak{a}_{\mathbf{Ht}}$	-5.506137	0.3931865	-14.00388	0
$\mathfrak{b}_{\mathbf{Ht}}$	2.341282	0.0789146	29.66854	0
$\mathfrak{a}_{\mathbf{Wt}}$	3.097235	0.11614581	26.66678	0
$\mathfrak{b}_{\mathbf{Wt}}$	0.8045534	0.02984497	26.95776	0
$\mathfrak{a}_{\mathbf{BSA}}$	5.782347	0.01818777	317.925	0
$\mathfrak{b}_{\mathbf{BSA}}$	1.249588	0.03331858	37.5042	0
$\mathfrak{a}_{\mathbf{BMI}}$	2.846198	0.3803801	7.48251	0.00E+00
$\mathfrak{b}_{\mathbf{BMI}}$	1.095032	0.1224039	8.946055	0.00E+00

Table 4.4: The 4 allometric scaling coefficients, i.e., $\mathfrak{b}$ values, between TCV and Ht, Wt, BSA, and BMI were calculated from the library (N = 97). Influential outliers were identified and removed using Cook's distance (> 4/N). As with Table 4.3, the backwards transform was $\boldsymbol{a} \triangleq \mathbf{e}^{\mathfrak{a}}$, i.e., $\mathfrak{a} = \boldsymbol{ln}(\boldsymbol{a}) \neq \boldsymbol{a}$.

Allometric Exponent Scaling Signals of Geometric Metrics with TCV

Coefficients	*Postulated Relationship Value*	*With Influential Outliers*		*Without Influential Outliers*	
		Value	*SE*	*Value*	*SE*
$\mathfrak{b}_{Ht}$	**3.0**	2.26	0.08	2.34	0.08
$\mathfrak{b}_{Wt}$	**1.0**	0.83	0.03	0.80	0.03
$\mathfrak{b}_{BSA}$	**1.5**	1.25	0.03	1.25	0.03
$\mathfrak{b}_{BMI}$	**3.0**	1.34	0.17	1.10	0.12

Table 4.5: The presented 4 allometric scaling exponent values, $\mathfrak{b}$, and their SEs were compiled from Tables 4.5 and 4.5. Wt and BSA's lower SEs suggested these models' coefficient estimates were relatively robust. The BMI and Ht SEs were large and mildly large to suggest these models' coefficient estimates were of relatively poor and slightly decreased robustness, respectively; Table 4.6's correlation values further supported this finding.

Healthy Heart Library Pearson's Correlations

	TCV	Age	Ht	Wt	BSA	BMI
TCV	1.000	0.763	0.846	0.901	0.920	0.670
Age		1.000	0.861	0.789	0.858	0.563
Ht			1.000	0.803	0.912	0.467
Wt				1.000	0.975	0.869
BSA					1.000	0.772
BMI						1.000

Table 4.6: Pearson's correlations for predictors were presented (N = 97). Sex and imaging modality were binary variables and therefore were excluded. Ht, Wt, and BSA had relatively strong correlation with TCV while the Age and BMI correlations with TCV were weaker.

4.3 Discussion of the Healthy Heart Library

The normal, healthy heart library was built to develop a novel virtual HTx fit assessment tool. Although the tool was originally geared towards the general HTx assessment; given:

1. the higher waitlist mortality rate in the youngest of patients,

2. the tool's concept was conceived to expand donor pools for patients with major cardiothoracic anatomy and physiology aberrations (largely in the CHD patient population), and

3. the clinical collaborators were pediatric cardiologist from a pediatric hospital,

the focus of the tool's development was geared towards pediatric HTxs. Specifically, for this thesis, it was perceived as being an unnecessary burden to include a larger adult, healthy heart population given the younger pediatric HTx population was most likely to benefit from the tool. The challenge in adding a larger adult healthy heart population and including adult transplant cases was (1) new collaborations would have needed to be made and (2) this would have expanded this work to a multicenter study. Keeping this work as a single center study resulted in only 5% of the healthy heart library being comprised of adults, i.e., 21+ years of age, with the oldest individual in the dataset being 29 years old.

Assessment of the Age, Ht, Wt, BMI, and BSA parameter results, as was shown in Table 4.2 and Figures 4.3 and 4.4, demonstrated the library encompassed a wide range of developmental growth, body types (e.g., overweight), and sizes present in the pediatric population. Requiring the library to encompass a large pediatric population range, which included a young adult population, helped to ensure the allograft TCV model makes interpolated predictions while avoiding extrapolated predictions. For example, including the available adult population in the library helped prevent extrapolated allograft TCV predictions in older pediatric donors (chapter 5). Unrealistic predictions can result when the model is made to extrapolate a prediction outside the training dataset range [108]. Less robust regression models, including models suffering from multicollinearity, are particularly at risk for extrapolating unrealistic predictions because their

estimated coefficients are unstable due to having large SEs - even when the extrapolation is mild [108,109].

The Sex demographic was biased towards male (64%) in the healthy heart library. In general, a data bias is not ideal from a regression modeling point-of-view (chapter 5). The male bias was similar in a practical sense to both the general pediatric HTx (56%) and the CHD pediatric HTx (61%) population biases [6,111,112]. Furthermore, no other health heart individuals were available at our collaborating institute within the IRB approved 15-year data window to increase the library's female population. Going outside this approved time duration to include more females would have changed this body of work to a prospective study – this would have increased the overall approval and execution challenges. Additionally, adding only prospective females to correct the male bias is not ideal from a statistical point-of-view because it could introduce nuisance factors (e.g., imaging protocol changes, types of healthy patients that were imaged, etc.). Given no other health heart individuals were available within the 15-year time-window and the male bias conformed reasonably well with the other important HTx populations, the Sex bias was considered no further and the library was moved forward in the tool's development process.

The one-way ANOVA results suggested the mTCV values were constant between modalities (p-value = 0.5156, Cohen's D effect size = 0.1328). The large p-value (>>> 0.05) suggested there was no statistically-significance difference between the CT and MR measurements. Specifically, from a statistical point-of-view, the null hypothesis that the mTCVs sample populations were the same was not rejected because the p-value was large. Furthermore, the effect size was small and corresponds to the CT and MR populations sharing a distribution overlap that was > 85% [113]. It is worth noting that even if the p-value was small, the small effect size, i.e., the large overlap between the two modality populations, and the small difference between the modality means, i.e., 34mL, would suggest the statistical-significance was meaningless in a practical sense. These results justify the CT and MR measurements were practically the same. The ANOVA test results support the stance that the CT and MR data can be combined in the healthy heart library without further consideration with what modality the data came from.

The estimated allometric proportional constants and scaling exponents between TCV and Ht, Wt, and BSA were relatively consistent with (Table 4.4) and without (Table 4.3) influential outliers removed. The consistency between the estimated coefficients was expected because their SEs were small, even when the influential outliers were kept. BMI was unusual for the 4 geometric parameters being fitted because the SEs were relatively large and resulted in a relatively large fluctuation between the coefficient estimates. The Pearson's coefficients in Table 4.6 demonstrated Ht, Wt, and BSA all correlated well with TCV, i.e., all 3 correlations were > 0.84 with Ht having the lowest correlation of the 3. BMI had a relatively poor correlation of 0.67. The relative variations in the SE for BMI were most probably a direct consequence of the correlation results, as demonstrated in Table 4.5. Age had a correlation of 0.76 and suggested the SE of a univariable fit with TCV might be between the BMI and the Ht, Wt, and BSA SEs. The unusual BMI SE and correlation results preluded an upcoming discussion in this section that BMI was not an indicator of developmental growth.

The scaling exponents were more closely analyzed in Table 4.5 and compared to their postulated, i.e., theoretical, scaling exponents with respect to TCV developmental growth. The key observation was the scaling exponents for Ht, Wt, and BSA had relatively minor deviations from their postulated values; the deviations were all hypoallometric in nature. BMI had a strong deviation from its postulated scaling exponent; the deviation was hypoallometric in nature. The correlation and amount of deviation from the postulated scaling exponents suggested Ht, Wt, and BSA described a major allometric scaling factor in developmental heart size, i.e., developmental growth, but not BMI.

Increase in Ht and BSA as an indication for pediatric growth was trivial. Increase in Wt as an indication for pediatric growth was less trivial because of body types, e.g., obesity. For example, in looking at the extreme ranges for Age and Wt in the healthy heart library we would never expect to have a 3-month-old weighing 139kgs or a 29-year-old weighing 3kgs. Matching these extreme and opposing Age and Wt ranges was perceived as clinically impossible for normal, healthy individuals and therefore supported Wt reflects developmental growth more than body type in the pediatric population. Similarly, the extreme TCV sizes differences between infants and adults

further supported Age helped to reflect developmental growth. It was trivial Age does not capture developmental growth between adults as growth generally stops after adolescents. Furthermore, the strong Pearson's correlations between Ht, Wt, and BSA developmental growth with TCV growth had a strong causality because as an infant develops into an adult it was expected the heart will need to grow much larger to produce sufficient cardiac output. The fact that the resting infant heart rate decreases from approximately 140 to 70 beats per minute in the adult further supports the heart must grow to produce sufficient cardiac output as the individual grows [114,115].

BMI was recognized as a metric of body type (e.g., overweight) and generally independent of Age, as visual inspection of Figure 4.4 suggested. The previous discussion that the TCV range in the library was more a reflection of developmental growth and the poor Pearson's coefficient reported between TCV and BMI further appears to have supported the BMI parameter does not reflect developmental growth. The only Age-related trend visually seen with BMI was that children and infants hold a constant (or near constant) BMI value, as was shown in Figure 4.4. This trend with Age and BMI might suggest infants and children "eat-to-grow" and after early developmental growth ends then poor, excessive eating habits (and lack of exercise) can result in overweight and obesity. This trend in children and infants having a small, relatively consistent body type, i.e., small BMI, might explain why the Pearson's coefficient was not closer to 0. The results suggested BMI does not indicate developmental growth but rather body type. A regression model might include BMI to help account for the effects, i.e., nuisance factors, of body type hidden within developmental growth (or possibly Wt with another developmental growth parameter).

From the TCV point-of-view, the TCV range (36 to 1340mLs) in the healthy heart library was large. Age, Ht, Wt, and BSA generally increased with respect to TCV. The strong Pearson's correlations between Age, Ht, Wt, and BSA and TCV (Table 4.6) supported the inference that the major factor contributing to heart size was developmental growth. The results highly suggested the allograft TCV prediction model developed in chapter 5 would largely be driven by one or more of the developmental growth components, i.e., Age, Ht, Wt, and BSA. The weaker correlation between Age and TCV suggested Age would not be the only developmental growth predictor if it were present in the final model. Finding Age had a weaker correlation with developmental growth is trivial

once it is recognized the oldest of the healthy heart subjects had likely stop growing, i.e., had finished going through puberty, for a period of time. Furthermore, the author was not aware of a causality for BMI being a driving force in developmental growth patterns but rather it was perceived as an indication for accounting for patient body type. The weak but measureable correlation between BMI and TCV suggested allograft TCV predictions might need to account for body type by adjusting the predicted TCV. As an analogues to BMI adjusting for body type, Sex was investigated in chapter 5 because the final model might need to adjust allograft TCVs based on whether the donor was male or female.

The univariable fits between TCV and the geometric predictors suggested allometric scaling were present in the data and suggested the final allograft prediction model would derive from these field of allometry concepts. In analyzing the geometric signal patterns, an interesting pattern in the scaling exponent (Table 4.5) was observed. In dividing the empirical scaling exponents (with influential outliers included) by the corresponding theoretical scaling exponents the signals' ratios were: 0.75 (Ht), 0.83 (Wt), 0.83 (BSA), and 0.44 (BMI). When the influential outliers were removed, the signals' ratios were: 0.78 (Ht), 0.80 (Wt), 0.83 (BSA), and 0.37 (BMI). It appears that parameters related to TCV developmental growth (i.e., Ht, Wt, and BSA) have an empirical to theoretical signaling ratio of approximately 0.80. BMI did not show this trend. Although this empirical-theoretical signal ratio pattern of $4/5^{ths}$ in TCV developmental could be happenstance, it was an unexpected trend outside the scope of this body of work. Future work might want to investigate if there is a causality for this biological signal and if it holds for other developmental growth signals.

Limitations of the healthy heart library, with regards to the novel virtual HTx tool's development, fell into two major categories. The limitation categories were: (1) the patients that were used as "healthy" subjects in the healthy heart library and (2) the collaborating hospital's clinical imaging practices.

First, the library was comprised of perceived normal, healthy heart patients. Given this was a retrospective study the "healthy" subjects were patients that received medical imaging because of a perceived clinical concern. Subjects ranged from trauma patients (in which cardiac anomalies

were not expected), to cancer patients (medications excluded a subset of these patients), to even cardiac patients (with minor cardiac aberrations). As the methods indicate, extreme retrospective subject vetting isolated only 97 subjects that could be used in the library out of thousands of patients that had CT or MR scans over a 15-year period of time. Close inspection of the 97 subjects was performed during data collection to help ensure the TCVs were normal, well-functioning organs.

Second, this retrospective study was dependent on what field-of-view images were archived and what imaging protocols were implemented. Only the archived images were available for TCV reconstruction because the original, i.e., raw, data was deleted to make room for new scans on the imaging systems. Furthermore, technicians would often truncate the apex of the heart while setting the final FOV to be archived because it was of no clinical value for the specific clinical case. The author was made aware of this archiving truncation practice, in which only the clinically relevant FOV is kept, through informal discussions with the scanner technicians.

Peak-diastole images, i.e., images acquired approximately at the 80% R-R interval of the electrocardiogram signal, were perceived as preferable because the ventricles would be at their largest volumes. Time-specific cardiac acquisitions are achieved through cardiac-gating, i.e., image acquisition was triggered by the electrocardiogram signal. Imaging the ventricles at their largest was perceived as important because the ventricles (blood volume and muscle mass) contribute a significant portion on the overall TCV. An additional benefit of imaging the heart at the 80% R-R interval was it corresponds to the longest, most static time-point in the cardiac cycle. Imaging the heart when it is at its most static time-point in the cardiac cycle helps to reduce risk of motion artifacts.

Hospital scan protocols and practices appeared to become more selective with what time-points were imaged over the 15-year period. In regard to the CT images, it was observed that obtaining images at the 80% R-R interval became more and more difficult as scan dates became more resent in the hospital's acquisition timeline. This appears to be due to a clinical shift from implementing retrospective to prospective CT acquisition protocols over the years. A logical causality in this observed trend might be radiologist wanting to implement prospective studies to reduce patient radiation exposure [116]. Many of the more recent prospective studies focused at

or around the 40% R-R interval, i.e., peak systole. MR imaging did generally keep retrospective scanning protocols in place with cardiac MR scans typically captured images at the 0%, 20%, 40%, 60%, and 80% R-R intervals. The lack of a shift away from MR retrospective scan protocols while there was a shift of the CT protocols helps to further support procedural changes were likely driven by patient radiation exposure concerns associated with CT scans.

Another issue with cardiac-gating was this method was not always implemented – this was often seen in trauma and lung-related cases. The consequences of not using cardiac-gating were (1) time-point specific cardiac images could not be acquired and (2) images would be slightly blurry around the TCV boundary due to motion artifact.

Due to the low healthy heart sample size, both images that were acquired outside the 80% interval and without cardiac-gating were included so an allograft TCV prediction model could be developed. However, the 80% R-R interval images or at least the cardiac-gated images were used when available.

Clinical MR CT scan quality for anatomical reconstructions were often limited by poor out-of-plane spatial resolution, driven by trade-offs addressing the poor temporal resolution of the MR modality [116,117]. Specifically, the source of poor spatial resolution was radiology tried to reduce scan times by reducing the number of slices they acquire within the heart's FOV [116,117]. Radiologist make the out-of-plane spatial resolution trade-off such that important anatomical and functional data is able to be acquired within the typical time-window clinicians clinically have with the patient. Many of the MR cardiac images had poor out-of-plane spatial resolution but good in-plane resolution. CT scanners have phenomenal temporal resolution and therefore it was rare to find CT scans with poor spatial resolution [116]. The ANOVA test, comparing CT and MR measurements and illustrated in Figure 4.5, helped justify the resolution differences did not affect the TCV measurements.

The health heart library was carefully pulled together to balance the realistic heart data that was available for this thesis body of work while acquiring the data needed to develop the novel virtual heart transplant fit assessment tool. Statistical analysis of the library justified the combination of CT and MR data and suggested the allograft prediction modeling process would be successful

if it included for field of allometry concepts. The library data had a large distribution of subjects to help ensure the developed model avoids both mild extrapolations and extrapolations. Although the library is of a low, moderate size, the data sample size met statistician recommendations to build the parametric, linear allograft TCV prediction model.

CHAPTER 5

DEVELOPMENT OF THE INITIAL AND IMPROVED ALLOGRAFT TCV PREDICTION MODELS

A regression model to predict normal, healthy allograft TCVs was developed as a part of the novel, virtual HTx fit assessment tool. This predictive model was developed in recognition that most donor CT or MR images would not exist or, if they do exist, they would not likely be readily available to the recipient's institution. Performing a virtual fit assessment with existing donor images is technically challenging because the donor images need to (1) be electronically transferred through a secured system, (2) have the donor TCV reconstructed, and (3) have the virtual assessment performed within the 1-hour time-window clinicians typically have to make a provisional acceptance.

The developed predictive model allows a clinical team to identify a TCV reconstruction in the healthy heart library (with a similar volume) to serve as an analogue for the actual allograft's geometry during the virtual fit assessment. For clinical utility, the model was developed to only consider gross donor metrics that were readily available at allograft offer with current clinical practices. To make use of the model, the novel tool needed to assume that (1) the donor cardiac allografts were normal, healthy, and well-functioning hearts and (2) TCVs of equivalent sizes had similar anatomical geometries. The predictive model allowed the tool to perform an assessment within 10 minutes while assessments driven by donor images required up to 30 minutes after images were transferred. Even with donor image assessments taking 30 minutes, it was recognized that transferring the donor CT or MR images between centers would be the major source for the time-related limitations because it is an unusual request. The perceived time-related data transfer bottlenecks included no current mechanisms set up to ease either the CT or MR acquisition of a deceased individual nor the expeditious transfer of the acquired data. The inclusion of the predictive model simplified the assessment process and increased the practicality of implementing the tool in the current clinical environment.

Chapter 5 covers the development and the validation of the allograft TCV prediction model. The model was developed from the health heart library dataset. The first half of chapter 5 covers the development and validation of the initial model, i.e., Model-A. The second half of chapter 5 covers improvements to the initial modeling process. These improvements produced the final model, i.e. Model-B, that were used in the virtual HTx fit assessment tool. Using Model-B's structural framework, the chapter includes the development of a more conservative third model, i.e., Model-B*. This conservative model was developed to limit, i.e., offset, the likelihood of allograft TCV under-predictions. Biasing Model-B to avoid under-predictions by favoring over-predictions was undertaken to avoid foreseeable complications when the tool is used to maximize the allograft a clinical team is willing to accept. For this body of work, the development and validation of the allograft TCV prediction model constituted as Aim 2.

5.1 Methods and Materials of the Allograft TCV prediction Model

All subjects in the healthy heart library (N = 97) were used to develop and validate the final allograft TCV prediction model from gross donor parameters. The modeling process considered and implemented only parametric, linear regression techniques - partly due to the limited population size of the healthy heart library [118]. Six gross parameters, listed in Table 5.1, were initially considered to be potential predictor variables in the prediction model. In chapter 4, an ANOVA test compared the CT and MR mTCV values (p-value = 0.5156, Cohen's D effect size = 0.1328). The ANOVA test results helped to justify that the current modeling process in chapter 5 did not need to account for modality type.

Considered Regression Modeling Variables

Predictor Variables:	***Classification or Units***
Sex	Male/Female
Age	months
Height (Ht)	cm
Weight (Wt)	kg
Body Surface Area (BSA)	m^2
Body Mass Index (BMI)	kg/m^2
Criterion Variable:	
Total Cardiac Volume (TCV)	mL

Table 5.1: The list of healthy heart library data that was pulled from patient charts and used to develop the allograft TCV prediction model. BSA and BMI were calculated from the pulled data.

The modeling process, i.e., model development and validation, consisted of 7 major steps as shown in Figure 5.1. Given (1) modeling steps 2 to 6 were repetitive in nature and (2) an exhaustive modeling search to identify top structural frameworks was implemented, it was sensible to semi-automate the modeling process. A model's "structural framework" refers herein to a mathematical equation comprised of a unique combination of predictor main effect and interaction terms while the coefficient values were left undetermined. The modeling process was semi-automated in a series of *R* language (R Core Team, Vienna, Austria) computer scripts [118,119]. *R* was chosen for its vast collection of open-source statistical packages that have been generated by a currently active statistical community. Implementing the available package functions helped to reduce the coding efforts of the author. The breaks between the series of in-house computer scripts allowed the author to evaluate and, if appropriate, implement minor modifications without rerunning the entire modeling process. Not needing to rerun the script from the beginning helped to reduce the modeling process's time duration because the exhaustive search was a computationally expensive procedure. The final allograft TCV prediction model was identified in step 7.

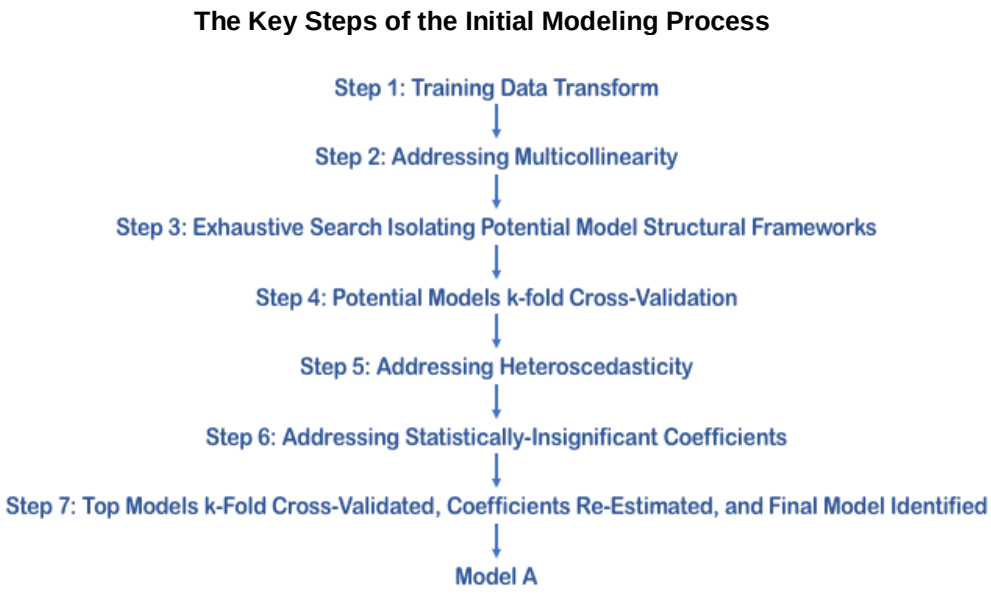

Figure 5.1: Diagram of the key steps used in the development of Model A. Cross-validation was implemented twice – midway through the modeling process and at the end.

Several statistical metrics were repeatedly evaluated throughout the modeling procedure. The mean absolute error (MAE), mean absolute percent error (MAPE), mean error (ME), mean percent error (MPE), mean square error (MSE), and root mean square error (RMSE) were a particularly important set of 6 statistical metrics used throughout this body of work. These 6 metrics quantitatively highlight unique attributes of the prediction error, i.e., residual. Consequently, this grouping of 6 metrics was henceforth referred to as the "statistical modeling metrics" within this thesis. It is noteworthy to highlight the importance in distinguishing errors derived from training and testing data (discussed later); therefore, "training statistical model metrics" and "testing statistical model metrics" were phrases used herein to further differentiate the type of "statistical modeling metrics" being analyzed. The corrected Akaike Information Criterion (AICc), symmetric mean absolute percent error (sMAPE), and symmetric mean percent error (sMPE) were additional metrics reviewed in the modeling process. These 3-additional quantities are statistical modeling metrics in their own rights but this thesis reserves this "statistical modeling metrics" terminology for the 6-aforementioned metrics. The reasoning for separating the 3-additional metrics was the AICc metric

is of unique importance and the remaining 2 metrics were not routinely referred to during the modeling process. All statistical metric formulas used herein were presented in Table 5.2.

The Formulas used to Quantify the Statistical Metrics

$ME = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i - \hat{y}_i}{1}$	$MAE = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{1}\right\rvert$
$MPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{y_i - \hat{y}_i}{y_i}$	$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
$sMPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{y_i - \hat{y}_i}{\frac{y_i+\hat{y}_i}{2}}$	$sMAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{\frac{y_i+\hat{y}_i}{2}}\right\rvert$
$MSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	$RMSE = \sqrt{MSE}$
$AIC_j = 2k_j - 2ln(\hat{L}_j)$	$AICc_j = AIC_j + \frac{2k_j(k_j+1)}{N-k_j-1}$

Table 5.2: The modeling metric formulas were combined within the current table. Notice the formulas within the right column were slight modifications of the corresponding equations in the left column. The formula variables were used as follows: y_i (actual TCV value), $\hat{y}_i$ (predicted TCV value), N (sample size, e.g., $N = 97$, unless Cook's distance suggested otherwise or k-fold cross-validation was being implemented), k_j (number of terms within a unique structural framework), and $\hat{L}_j$ (estimated maximum likelihood). In general, the indices "i" and "j" corresponded to a specific data point and to a specific statistical model, respectively. During the 100 k-fold cross-validation runs, "i" represented a data point within a specific fold.

Step one, the models were fitted using non-transformed, i.e., original-value, and transformed (henceforth isometric* and "allometric*) training datasets. The transforms were 2 methodical attempts to linearize the original data such that applying linear regression techniques would allow for field of allometry concepts to be implemented in the substructure of the models' structural frameworks. The isometric* and allometric* refer to the values being transformed such that specific field of allometry concepts could be considered. The use of isometric* data and

allometric* data does not signify the transforms made the datasets isometric or allometric in nature (as this would be self-defeating for the modeling process). These 3 non-transformed and transformed datasets were referred herein as the 3 “datatypes”. Considering these 3 different datatypes in the modeling process was one of the fundamental causations for modeling steps 2 to 6 to be repetitive in nature.

The “isometric” data herein was previously transformed such that a linear regression modeling process would generally fit the data well if isometric scaling assumptions held between the predictors and the TCV. The predictor variables were transformed by applying appropriate powers to scale the variables’ characteristic units to that of volume. For example, to match height (m) and BSA (m^2) to the volume’s characteristic unit (i.e., m^3), their values were transformed by taking the powers of 3 and 1.5, respectively. In other words, power-law functions were carefully crafted such that the independent variables would be transformed to force an investigation of the isometric assumption. Furthermore, the density of the entire, i.e., whole, human body can generally be accepted as a constant that is near equivalent to soft tissue in an individual with a normal, healthy body weight [120,121]. The body density constant implies the body weight’s characteristic unit is equivalent to volume; therefore, in this work, weight was taken to the power of 1 to force an investigation of the isometric assumption. The 1:1:1 scaling of body weight, TCM, and TCV and the field of allometry concepts, both presented in chapter 2, further support scaling weight by the power of 1. The Age and Sex predictors were not transformed (yet still considered in the model) because they were non-geometric variables. The set of transforms applied to investigate an isometric scaling relationship were henceforth combined together and were referred to as the “cubed transform” – “cubed” is referring to the characteristic unit of volume.

The “allometric” data herein does not require an isometric scaling for a linear regression to fit the data well; however, isometric scaling relationships could be discovered in the process. Allometric scaling was considered in the modeling process by *log* transforming the data. The *log* transforms used *log* based *e*, i.e., *natural log*. Unlike the cube transform, the *natural log* transforms were not specifically tailored to geometric scaling relationships between a predictor and the characteristic unit of TCV; therefore, Age was *log* transformed. Although Sex could also be *log*

transformed, it is a nominal, categorical variable that, if transformed, would (1) not add to the modeling process, (2) complicate the interpretation of the model, and (3) be problematic because it is commonplace to use the dummy-variable “0”. Specifically, the resultant of *log* transforming “0”, i.e., $log(0)$, is undefined (and its right-sided limit goes to negative infinity). Therefore, Sex was not transformed but it was considered in the model. Henceforth, the set of transforms applied to investigate a generalized allometric scaling relationship were combined together and were referred to as the “*log-log* transform” – “*log-log*” is referring to both the dependent and independent variables being *log* transformed.

Step two, multicollinearity between predictor variables were tested for and appropriately addressed in an iterative process for each of the 3 datatypes. The variance inflation factor (VIF) was used to test for collinearity. Iteratively the predictor with the largest VIF value was systematically removed and the remaining predictors reanalyzing until the largest VIF value achieved a threshold < 10. A VIF of 1 indicates no collinearity between two variables while a value greater than or equal to 5 or 10 generally indicates strong collinearity [108,109,118]. Removing predictors until the VIF condition was met, helped to ensure the coefficient estimates were robust and generally meaningful with respect to the dependent variable [108,109,118]. As will be seen later with the improved modeling process, the iterative removal of the largest VIF first is not a set procedural requirement for addressing collinearity issues.

A generalized, i.e., reference, modeling procedure was included herein, where no collinearity corrections were implemented. This reference model was included to demonstrate the importance of addressing collinearity issues in the modeling process. Although high collinearity does not necessarily degrade the predictive performance of a modeling process (especially when training errors are being analyzed to assess the performance), it is well known that collinearity issues reduce the robustness of the produced models [108,109,118]. Specifically, skipping the VIF procedure put the reference modeling process at risk by allowing the exhaustive search procedure, i.e., step 3, to consider models with collinearity issues for candidacy in the group of “top models”.

Step three, to identify the potential structural frameworks for the final allograft TCV prediction model, an exhaustive search was performed. The exhaustive search was implement

using the *R* “glmulti” package [122,123]. Main effects (predictors) and pair-wise interactions (multiplication of two predictors) were considered terms within the structural frameworks. Three-way and larger interactions were not considered due to (1) a limitation in the “glmulti” code and (2) the computational cost required to include larger interactions [122]. The AICc metric was used, instead of the classical R^2 and adjusted-R^2 metrics, to quickly compare the exhaustive search structural framework performances in which the number of included terms varied for each model’s structural framework [118,122–124]. For comparative purposes, the AICc metric requires training datasets of the same size or, even better, the same dataset; therefore, all 97 healthy heart library data points were used in each of the exhaustive search fits [108,118,125,126]. Structural frameworks with lower AICc values were generally perceived to be better predictive models [124].

Using the lowest AICc value as a reference, AICc differences for all the exhaustive search results were determined. Literature suggest that AICc difference criteria of > 10, > 4, and ≤ 2 have negligible, some, and substantial chance that the two models being compared have similar performances; however, choosing the final threshold criterion one uses is somewhat subjective [124]. An AICc difference threshold of ≤ 2 is a typical, i.e., historical, guideline to help reduce the number of model structural frameworks needed to be considered in the modeling process [122,124–128]. Using this historical guideline, models with AICc differences of ≤ 2 were moved along in the modeling process. The small AICc difference criterion provided strong evidence that the models moved along in the modeling process had relatively equivalent prediction performances and therefore a close investigation of their testing statistical metrics was required to tease out the “best” model.

Step four, the potential models underwent a k-fold ($k = 10$) cross-validation process to identify, e.g., help validate, well performing structural frameworks that were not sensitive to specific training and testing datasets [118]. Cross-validation procedures validate models by systematically separating the data into a series of training and testing data subsets at random to estimate both the model’s coefficient(s) and the testing error several times. Estimating the testing error several times further allows for the testing error variance to be estimated. Ultimately, this process helps to ensure the model is robust to both the training and testing datasets that are used. Cross-validation

procedures are particularly useful in validating model structural frameworks when the data available for both the training and testing datasets are of limited size [118]. The healthy heart library (N = 97) was of moderate, i.e., limited, size and therefore supported the need for implementing a cross-validation procedure.

The k-fold procedure was run 100 times to remove bias in a single k-fold procedure. Running a 10-fold procedure 100 times corresponded to developing and testing the model 1000 times. The 100 runs were achieved in the semi-automatic script by using 100 unique seeds to randomly generate the k-fold subsets. These seeds allow for the same "random" subsets to be algorithmically regenerated to ultimately ease result comparisons between the different models. Furthermore, these 1000 folds produced 1000 testing statistical model metric results that were averaged to remove the aforementioned fold bias. The averaged testing statistical model metric results were then used to identify up to the top 10 potential model structural frameworks per datatype (up to 30 models in total) that were to move forward in the modeling process. In addition to averaging the 1000-fold testing statistical model metric results, their minimum, 1^{st}-quartile (1^{st}-Qt), 2^{nd}-quartile (2^{nd}-Qt ; i.e., median), 3^{rd}-quartile (3^{rd}-Qt), maximum, and standard deviation (St. Dev.) were also determined. It is worth noting this same k-fold cross-validation procedure was repeated a second time, in step 7, to select the final, "best" model.

Step five, heteroscedasticity in the top models were tested for and corrected when needed. Visual inspections of Figures 4.2 and 4.3 initially suggested heteroscedasticity in the library training dataset. Residual graphical visualizations and a heteroscedasticity test, i.e., the Breursch-Pagan test, were used to confirm heteroscedasticity [129,130].

Two methods to correct for heteroscedasticity were considered. The first method used the "varclass" functions in the *R* "nlme" package [129,131]. The "varclass" function uses sub-functions to reweight the coefficient estimates by applying mathematical assumptions to the data's variance. For example, the "varExp" sub-function generates weights that correct for heteroscedasticity by assuming the variance has an exponential fit between a dependent and independent variable paring. The various sub-functions were considered for each of the independent and dependent variable pairings in an iterative fashion. Heteroscedasticity was indirectly suggested when applying

"varclass" functions improved the AICc and the training statistical model metric results. The second method used Cook's distance criterion (4/N) to identify and remove influential outliers from the training dataset [108,109]. It is worth nothing the 4/N Cook's distance cutoff is often presented in literature as $2/\sqrt{N}$. The AICc and the training statistical model metric results produced by both methods were analyzed to determine which, if any, corrections were needed to reduce heteroscedasticity issues in the modeling process.

Step six, the near final estimated coefficient p-values were assessed to ensure all model terms were statistically-significant. Estimated coefficients with p-values of ≥ 0.05 and ≥ 0.025 were considered to be automatically and potentially insignificant, respectively. Statistically-insignificant coefficients can likely be removed from the model because they should have negligible effect on the remaining estimated coefficient values and the overall prediction error (excluding the prediction issues associated with an overfitted model) [108,109,118]. If coefficient values were candidates for removal, in an iterative fashion, the model was re-estimated with the largest p-value coefficient term removed and the training statistical metric values recalculated. Remaining coefficient estimates, SEs, and training statistical metric values were monitored throughout the insignificant coefficient removal process to ensure minimal changes were observed, i.e., to ensure no negative impact was observed in a term's removal. This iterative process was repeated until all coefficients with p-values ≥ 0.025 were removed without negatively affecting the model. Given the previously implemented modeling steps, the author was advised that large coefficient estimate p-values (≥ 0.05) should not be expected in step 6. A large coefficient p-value would likely have signified a serious issue in the design and/or implementation of the modeling protocol herein and therefore would need to be investigated and corrected.

It is worth mentioning that both steps 5 and 6 used all N = 97 data points to quantify the errors and, in general, used these same data points to train these same potential models (removal of influential outliers by using Cook's distance was the exception). No data holdout to calculate the testing error was performed. Given the reported errors in these 2 steps were comprised entirely or largely of training errors (with influential outliers included) the errors in these steps were referred to as training errors herein. Analyzing the training errors in steps 5 and 6 was considered acceptable

because (1) it simplified the heteroscedasticity and the insignificant coefficient correction processes and (2) no model selection (outside of model permutations) was performed in these steps. In particular, the modeling process was studiously crafted to ensure actual model structural framework selection steps were driven solely by testing error metrics in steps 4 and 7.

Step seven, the approximately top 10 models from each of the 3 datatypes (30 in total) were to be 10-fold cross-validated a second time. Additionally, each of the models were fitted one last time using all N = 97 data points in the final training dataset (excluding the influential outliers the Cook's distance procedure removed). The testing statistical model metric results from the 10-fold cross-validation were used to identify the final, "best" performing allograft TCV prediction model. The reported testing predicted TCV (pTCV) results were determined by averaging the 100 10-fold cross-validation runs. The design of step 7 allowed for the 'best' model to be selected from testing errors (by implementing a second k-fold cross-validation) and for the 'best' model to help reduce future allograft TCV prediction testing error overestimates (by estimating the final coefficients will all 97 data points).

The usage of "best" model, in step 7, was used to indicate (1) the modeling process was somewhat subjective, (2) the modeling process was dependent on model requirements and assumptions, and (3) the very label of best can be misleading between extremely similar performing models. The subjective nature refers to both the designer's selection of statistical modeling metrics to be analyzed and the designer's overall preference for one metric over another. A large variety of statistical metrics were included herein to ensure various error attributes were considered in the modeling process and to avoid biasing a specific error attribute. Furthermore, the designer requirements and assumptions herein were carefully considered based on concepts and/or practices from the fields of allometry, statistics, and medicine. Finally, the issue of labeling a model as "best" out of a set of similar performing models (without the quotations) is training and testing dataset variations might fluctuate in ranking performance order. However, the cross-validation procedures implemented in steps 4 and 7 helped to minimize the ranking order fluctuation issues between the training and testing samplings used at structural framework selection. The initial

modeling protocol presented herein was ultimately designed to help ensure a strong, final allograft TCV prediction model was developed – this initial model was referred herein as Model A.

5.2 Results of the Allograft TCV prediction Model

The 3 previously mentioned datatypes, transformed from the healthy heart library data in step 1, were considered for prediction model development. In step 2, collinearity issues for each of the 3 datatypes were addressed through a systematic elimination procedure using a VIF criterion threshold (≥ 10). The final predictors isolated for model consideration, from each of the datatypes, were as follows: original-value (Sex, Age, Ht, and BMI), *log-log* transform (Sex, Ht, and BMI), and cubed transform (Sex, Age, Ht, and BMI). Before the VIF procedural corrections were implemented, the initial maximum VIF values for each of the 3 datatypes were 820.60 (original-value; BSA), 302260.70 (*log-log* transform; Wt), and 20740.54 (cubed transformed; BSA). After the VIF procedural corrections were completely implemented, the final maximum VIF value for each of the 3 datatypes were 4.81 (original-value; Age), 1.32 (*log-log* transform; BMI), and 4.14 (cubed transformed; Age) - all other remaining predictors had VIF values less than the maximum VIF for their datatype. Interestingly, the VIF procedure for all 3 datatypes was constant in removing BSA and Wt in the first two iterations, i.e., if BSA was removed first then Wt was removed second.

Clinical considerations early within this body of work considered excluding Age automatically (see discussion); therefore, the VIF process was repeated with Age removed. For all 3 datatypes, with Age removed, the VIF procedure reduced the predictors to the same 3 final variables for consideration in the modeling process: Sex, Ht, and BMI. Regardless if Age was initially included or excluded in the modeling process, the *log-log* transform modeling process resulted in the VIF procedure isolating the same 3 predictors. This observation of the *log-log* transform modeling processes isolating the same 3 predictors was of particular importance because it helped contribute to a procedural simplification made after the completion of step 3. With Age excluded, the final maximum VIF values and their corresponding predictor for each of the 3

datatypes were 1.28 (original-value; Ht), 1.32 (*log-log* transform; BMI), and 1.16 (cubed transformed; Ht).

The isolated predictor sets with their corresponding non-transformed or transformed datatypes underwent an exhaustive modeling search procedure, in step 3, to identify potential model structural frameworks for allograft TCV prediction. The *R* "glmulti" function was used to implement the exhaustive search. All 97 data points, for each of the 3 datatypes, were used to train each of the exhaustive search models and calculate their corresponding AICc values. Keeping the data set constant, i.e., keeping all 97 data points, eased interpretation of the AICc results. Using the different transformed data for AICc comparison was appropriate because it only manipulated the data points – it did not change the specific individual's data points used.

Considering Age within the modeling process, the lowest AICc values in the exhaustive search were 1160.4 (original-value), -85.3 (*log-log* transform), and 1145.1 (cubed transformed). Automatically excluding Age within the modeling process, the lowest AICc values in the exhaustive search were 1170.8 (original-value), -85.3 (*log-log* transform), and 1146.2 (cubed transformed). The AICc values were the same for both *log-log* transform modeling processes (with and without Age considered) because they ended up being the exact same exhaustive search executions, i.e., these two exhaustive search procedures considered the same 3 predictors, used the same training dataset, and used the same datatype.

A reference exhaustive modeling search procedure, to demonstrate the importance of correcting for high collinearity issues, was also implemented with all 6 original predictors considered. The reference modeling process was achieved by skipping the VIF procedure. The lowest AICc values in the reference exhaustive search were 1141.9 (original-value), -88.2 (*log-log* transform), and 1133.8 (cubed transformed).

A very important observation with the AICc results was all 3 *log-log* transform modeling processes, i.e., with and without Age included and the reference modeling process, had excessively lower-starting, lowest AICc values (reported above) compared to the original-value and cubed transformed datasets. These intermediate results supported the *log-log* transform modeling processes well outperformed the other processes and therefore the original-value and cubed

transformed modeling processes were immediately terminated. Although the reference modeling process had slightly lower-starting, lowest AICc values than the VIF corrected models there was concern the reference models were not robust due to collinearity related issues. The AICc results provide no information on model robustness issues and could be hiding overfitting issues. Furthermore, the difference between the lowest AICc values produced by the non-reference and reference *log-log* transform fitted models was minimal, i.e., < 3. Both the concerns that the reference model was not robust and the observation the AICc difference was minimal, i.e., provided evidence these models' performances were nearly equivalent, justified a more in-depth analysis to tease out the "best".

As a consequence of both the VIF predictor reduction procedure results and the exhaustive search AICc results, the number of modeling processes needed to be considered was reduced to a single modeling process. This single modeling process needed to only consider *log-log* transform data with only 3 predictors, i.e., Sex, Ht, and BMI. As was previously shown, the inclusion or exclusion of Age had no effect on which predictors remained for consideration after the VIF procedure was implemented on the *log-log* transform data. Moving forward, due to the simplifications that could be made in the modeling process hereto, the allograft TCV prediction modeling process and reference modeling process were interchangeably referred to as the "VIF corrected" and the "non-VIF corrected" models, respectively.

The VIF corrected, *log-log* transform modeling procedure had 64 model structural frameworks to investigate in the exhaustive search. The AICc difference threshold criterion (≤ 2) isolated 8 (out of 64) structural frameworks to be moved forward in the modeling process, i.e., move forward to the k-fold cross-validation procedure in step 4. The 64 AICc results were illustrated in Figure 5.2 by plotting the AICc values against their sequential AICc ranking. The 8 potential models that were moved forward in the modeling process were marked red in Figure 5.2 and their structural frameworks were presented in Table 5.3. Based on the sequential AICc ranking order, these 8 models were referred to as potential Models 1, 2, …, and 8. A zoomed in view of the AICc ranking profile in Figure 5.2 illustrated a discontinuity between potential Models 6 and 7. Although a further investigation was needed to determine if Models 7 or 8 would have performance limitations, the

AICc profile discontinuity suggested neither models 7 or 8 would likely be selected as the final "best" model. For comparison, the reference modeling process isolated 74 (out of approximately 4096) structural frameworks that were moved forward by using the same AICc difference criterion (≤ 2). The number of reference model structural frameworks analyzed in the exhaustive search was approximated using the provided equation 2^{p^2} by the "glmulti" author – where "p" was 6, i.e., was the number of predictors [122]. The VIF corrected modeling process did not need to approximate the number of structure frameworks analyzed because the "glmulti" reports up to the lowest 100 AICc models and therefore the exact number was known.

AICc Profiles from Initial Modeling Process's Exhaustive Search

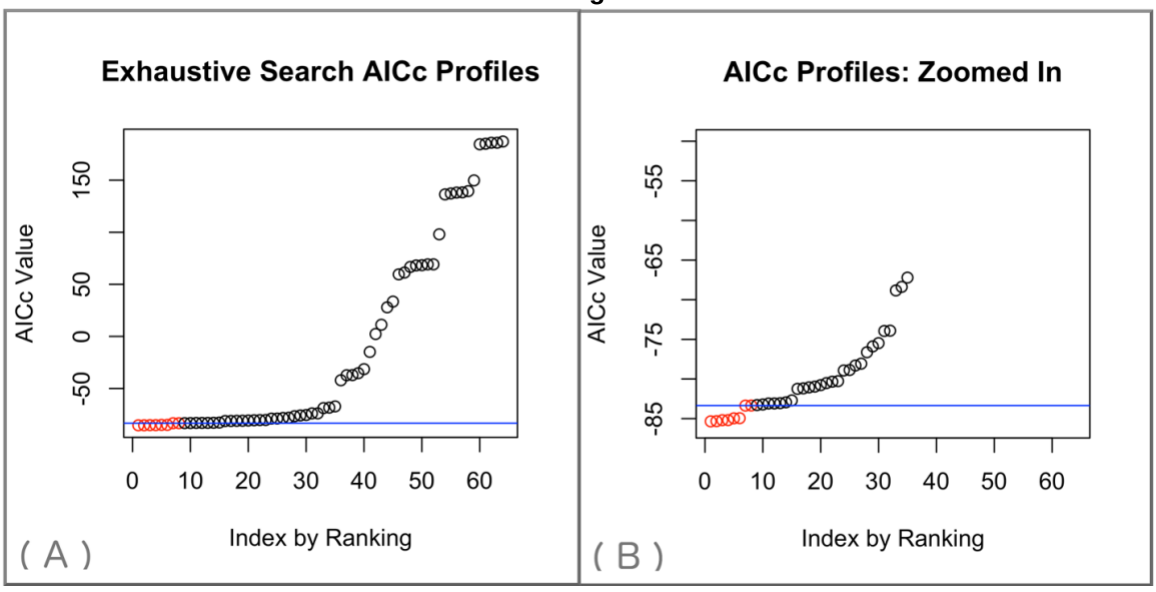

Figure 5.2: The 64 AICc profiles sequential order rankings were illustrated with the 8 potential models (marked red) below the AICc difference threshold (blue, horizontal line). Although image (A) suggested minimal patterns of AICc profile discontinuity in the first 32 models, image (B), zoomed in, illustrated several discontinuities. One discontinuity of particular interest was between potential Models 6 and 7. The AICc values are from the VIF corrected exhaustive search modeling procedure in which both the ML method and all healthy heart data points (N = 97) were used.

Potential Models Identified by AICc Difference Criteria

Model	*Structural Framework (R software notation)*	*AICc*
1	TCV ~ 1 + Ht + BMI + BMI:Sex	-85.35
2	TCV ~ 1 + Ht + BMI:Sex + BMI:Ht	-85.32
3	TCV ~ 1 + Ht + Ht:Sex + BMI:Ht	-85.20
4	TCV ~ 1 + Ht + BMI + Ht:Sex	-85.19
5	TCV ~ 1 + Sex + Ht + BMI:Ht	-84.96
6	TCV ~ 1 + Sex + Ht + BMI	-84.94
7	TCV ~ 1 + Sex + Ht + BMI + Ht:Sex	-83.36
8	TCV ~ 1 + Sex + Ht + Ht:Sex + BMI:Ht	-83.35

Table 5.3: The 8 potential structural frameworks, that were moved forward in the modeling process, were presented in AICc sequential order. These frameworks were identified in an exhaustive search in which only Sex, Ht, and BMI were considered in the modeling process. As suggested by the VIF correction process, Age, BSA, and Wt were excluded to prevent high multicollinearity within the independent variables. The AICc values were produced from a modeling procedure in which both the ML method and all healthy heart data points (N = 97) were used.

Before moving on to the cross-validation of these 8 potential models in step 4, an important clarification needs to be made on what AICc results were presented hereto and the AICc results presented henceforth. The “glmulti” function used the maximum likelihood (ML) to calculate the AICc values for model comparison [123]. However, the restricted ML (REML) is a more robust and preferred method of the ML technique for final model coefficient estimation [129]. When it comes to comparing the performance between two or more structural frameworks, by using their AICc values, there are well-established statistical reasons the ML, not the REML, needs to be used [129,132]. The concerns for using the REML method for model comparison very likely explain why the “glmulti” function did not have the REML as an option. Given these listed factors, the REML method was used herein for coefficient estimates analysis and final model reporting while the ML method was only used for comparisons made between two distinct structural frameworks. Nevertheless, REML derived AICc values were still reported herein as an easy method to suggest slight performance improvements gained for a specific structural framework through the implementation of heteroscedasticity corrections in step 5; however, these REML derived AICc

values were not used to make comparisons between distinctly different structural frameworks. In summary, the previous exhaustive search modeling AICc values (e.g., Figure 5.2 and Table 5.3), using ML, and the REML derived AICc values (to be presented shortly) were not comparable.

The 8 potential VIF corrected models were k-fold cross-validated in step 4 and their testing statistical model metric results were summarized in Tables 5.4-15. The reported results were a summarization for each of the testing statistical model metrics by determining each metric's minimum, 1^{st}-quartile, mean, 2^{nd}-quartile (i.e., median), 3^{rd}-quartile, maximum, and standard deviation for the 1000 folds. For example, Model 1's reported mean in Table 5.4, i.e., 65.6mL, was calculated by taking the mean of the 1000 MAE values determined for each of the 1000 folds. For Tables 5.4-15, the even tables specifically presented the results of the 8 potential models and the odd tables focused on making comparisons between the 8 potential modeling and the reference modeling results.

Tables 5.4, 5.6, 5.8, 5.10, 5.12, and 5.14 presented testing statistical model metric result summaries for each of the 8 potential models. In-depth cross-validation analysis for each of the 8 potential models did not identify any major testing error deviations between any of the model structural frameworks. The results suggested all 8 models were generally equivalent in prediction performance and robustness. The near-equivalent percentage errors, i.e., averaged MAPE (~ 12%) and MPE (~ -1%), and their standard deviations for all 8 models were perceived to be reasonable for the final model's intended application – even though these specific percentage metrics have a well-established asymmetrical, bias (see discussion). The slight negative MPE values indicated a slight over-prediction bias for all 8 potential models. These percentage errors were important to consider because, for example, a 50mL allograft prediction error was perceived to be less of a concern for a teenager (e.g., a native 800mL TCV heart) then it was for an infant (e.g., a native 75mL TCV heart). Nevertheless, there was a negligible but detectable trend that Model 2 might have performed "best" because it generally had the lowest (or near lowest) testing statistical model metric means, medians, and standard deviations. Model 2 specifically had the lowest means, medians, and standard deviations for the MSE and RMSE metrics – these two metrics are of

particular importance because they penalize larger errors (discussed later). Furthermore, analyze of the testing statistical model metrics suggested Model 3 was a close second in performance.

Tables 5.5, 5.7, 5.9, 5.11, 5.13, and 5.15 compared the performances of the VIF corrected and the reference modeling processes though a series of testing statistical model metric summaries. The VIF corrected and reference testing metric summaries were derived from the 8 and 74 potential models, respectively. In general, the lowest AICc reference model testing statistical model metric means and medians, for the 1000 folds, were closer to zero then for Model 1 (all metrics going to zero would indicate perfect predictions). Although the lowest AICc reference model produced better testing statistical model metric results, its corresponding testing statistical model metric variances were larger than Model 1's variances and therefore indicated the reference model was less robust. Furthermore, averaging the 8 VIF corrected and 74 reference testing statistical model metric means, medians, and standard deviations further suggested (1) the VIF corrected models were more robust (had lower variances) and (2) the VIF corrected models, on average, had equal or slightly better prediction performances. This reduced robustness of the models produced during the reference modeling process was likely a consequence of high collinearity in the data that was not corrected. These testing statistical model metric result variations were dependent on the developed models' robustness because the robustness determined how sensitive the results were to a specific testing and training dataset combination during a given fold.

No large deviations between the 8 VIF corrected potential models' prediction performance averages, i.e., testing statistical model metric results, in Tables 5.4-15, were observed in the cross-validation process. The minimal differences in these 8 models' performance averages could not justify preferential treatment or immediate termination of any models at this point in the modeling process. Model 2 and, to lesser extents, Models 3 and 5 did have slight trends suggesting they outperformed the other 5 models in prediction capability. However, trends indicated the variance of the VIF corrected models increased as the model number, i.e., Model 1, 2, …, 8, increased. Models 3 and 5, with respect to Model 2, did have larger variances in their prediction performances. The averaged performance ranking trends for Models 2, 3 and 5 held with respect to their MSE and RMSE metrics. Hereto, cross-validation results suggested Model 2 might be the final structural

framework used to predict allograft TCV but given (1) the minimal performance deviations between the models and (2) there were less than 10 models to consider, all 8 VIF corrected models were moved forward in the modeling process. Henceforth, these 8 models were further considered as top potential models in the modeling process.

Potential Models' Cross-Validated Mean Absolute Errors

Model	***Minimum***	***1st-Qt***	***Mean***	***2nd-Qt (Median)***	***3rd-Qt***	***Maximum***	***St. Dev.***
1	29.6mL	51.1mL	64.3mL	65.6mL	78.8mL	101.4mL	18.1mL
2	30.4mL	50.9mL	63.9mL	65.5mL	78.8mL	100.9mL	18.1mL
3	29.4mL	49.5mL	63.9mL	65.2mL	79.4mL	101.0mL	18.3mL
4	28.6mL	49.8mL	64.3mL	64.6mL	80.2mL	101.5mL	18.3mL
5	29.8mL	49.6mL	64.0mL	64.9mL	79.8mL	100.8mL	18.3mL
6	28.9mL	49.8mL	64.5mL	64.5mL	80.1mL	101.3mL	18.4mL
7	28.0mL	49.4mL	64.5mL	65.0mL	79.5mL	101.8mL	18.3mL
8	28.8mL	49.2mL	64.0mL	66.5mL	79.1mL	101.2mL	18.2mL
Mean	29.2mL	49.9mL	64.2mL	65.2mL	79.5mL	101.2mL	18.2mL
St. Dev.	0.8mL	0.7mL	0.3mL	0.7mL	0.6mL	0.3mL	0.1mL

Table 5.4: Cross-validation results suggested all 8 models have an averaged testing MAE of 64mL and an averaged standard deviation of 18mL. The low standard deviations between the 8 structural frameworks' testing MAE metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Mean Absolute Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	*Lowest AICc Model*	29.6mL	51.1mL	64.3mL	65.6mL	78.8mL	101.4mL	18.1mL
	Mean of All Potential Models	29.2mL	49.9mL	64.2mL	65.2mL	79.5mL	101.2mL	18.2mL
	St. Dev. of All Potential Models	0.8mL	0.7mL	0.3mL	0.7mL	0.6mL	0.3mL	0.1mL
All 6 Predictors Considered	*Lowest AICc Model*	29.4mL	50.0mL	63.7mL	64.9mL	75.6mL	107.6mL	19.6mL
	Mean of All Potential Models	30.7mL	49.8mL	64.9mL	66.4mL	77.9mL	107.7mL	19.5mL
	St. Dev. of All Potential Models	2.4mL	2.2mL	1.5mL	2.3mL	1.8mL	5.3mL	1.0mL

Table 5.5: The lowest AICc VIF corrected and reference models' cross-validation testing MAE means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

Potential Models' Cross-Validated Mean Absolute Percentage Errors

Model	*Minimum*	*1st-Qt*	*Mean*	*2nd-Qt (Median)*	*3rd-Qt*	*Maximum*	*St. Dev.*
1	5.88%	10.24%	12.07%	12.09%	13.69%	17.14%	2.45%
2	5.98%	10.27%	12.01%	12.00%	13.58%	17.13%	2.41%
3	5.82%	10.27%	12.00%	11.90%	13.70%	17.15%	2.46%
4	5.71%	10.24%	12.06%	12.01%	13.77%	17.16%	2.51%
5	5.83%	10.37%	12.03%	12.10%	13.70%	17.15%	2.46%
6	5.73%	10.35%	12.09%	12.15%	13.82%	17.16%	2.51%
7	5.83%	10.19%	12.17%	11.89%	14.66%	17.04%	2.57%
8	5.93%	10.13%	12.08%	11.77%	14.23%	17.03%	2.51%
Mean	5.84%	10.26%	12.06%	11.99%	13.89%	17.12%	2.48%
St. Dev.	0.09%	0.08%	0.05%	0.13%	0.37%	0.05%	0.05%

Table 5.6: Cross-validation results suggested all 8 models have an averaged testing MAPE of 12% and an averaged standard deviation of 2%. The low standard deviations between the 8 structural frameworks' testing MAPE metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Mean Absolute Percentage Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	*Lowest AICc Model*	5.88%	10.24%	12.07%	12.09%	13.69%	17.14%	2.45%
	Mean of All Potential Models	5.84%	10.26%	12.06%	11.99%	13.89%	17.12%	2.48%
	St. Dev. of All Potential Models	0.09%	0.08%	0.05%	0.13%	0.37%	0.05%	0.05%
All 6 Predictors Considered	*Lowest AICc Model*	6.81%	9.87%	11.91%	10.86%	13.25%	18.45%	2.90%
	Mean of All Potential Models	6.73%	10.03%	12.12%	11.38%	13.72%	18.81%	2.88%
	St. Dev. of All Potential Models	0.77%	0.30%	0.22%	0.38%	0.39%	1.30%	0.37%

Table 5.7: The lowest AICc VIF corrected and reference models' cross-validation testing MAPE means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

Potential Models' Cross-Validated Mean Errors

Model	*Minimum*	*1st-Qt*	*Mean*	*2nd-Qt (Median)*	*3rd-Qt*	*Maximum*	*St. Dev.*
1	-41.1mL	-10.6mL	8.2mL	3.6mL	33.8mL	57.3mL	26.3mL
2	-40.6mL	-10.7mL	8.0mL	4.7mL	32.7mL	58.0mL	26.0mL
3	-39.8mL	-11.0mL	8.0mL	5.8mL	32.0mL	57.4mL	25.5mL
4	-40.2mL	-10.9mL	8.3mL	4.7mL	33.3mL	56.6mL	25.8mL
5	-39.9mL	-10.6mL	8.1mL	5.6mL	31.9mL	57.5mL	25.5mL
6	-40.4mL	-10.5mL	8.3mL	4.6mL	33.2mL	56.7mL	25.7mL
7	-39.6mL	-12.1mL	8.2mL	5.7mL	33.7mL	56.9mL	26.1mL
8	-39.2mL	-12.1mL	7.9mL	6.8mL	32.3mL	57.6mL	25.8mL
Mean	-40.1mL	-11.1mL	8.1mL	5.2mL	32.9mL	57.2mL	25.8mL
St. Dev.	0.6mL	0.7mL	0.2mL	1.0mL	0.8mL	0.5mL	0.3mL

Table 5.8: Cross-validation results suggested all 8 models have an averaged testing ME of 8mL and an averaged standard deviation of 26mL. The low standard deviations between the 8 structural frameworks' testing ME metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Mean Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	***Lowest AICc Model***	-41.1mL	-10.6mL	8.2mL	3.6mL	33.8mL	57.3mL	26.3mL
	Mean of All Potential Models	-40.1mL	-11.1mL	8.1mL	5.2mL	32.9mL	57.2mL	25.8mL
	St. Dev. of All Potential Models	0.6mL	0.7mL	0.2mL	1.0mL	0.8mL	0.5mL	0.3mL
All 6 Predictors Considered	***Lowest AICc Model***	-38.6mL	-21.4mL	7.1mL	7.2mL	35.3mL	56.7mL	27.0mL
	Mean of All Potential Models	-41.3mL	-15.3mL	7.2mL	6.5mL	32.7mL	58.0mL	26.8mL
	St. Dev. of All Potential Models	3.0mL	5.7mL	0.8mL	1.6mL	2.9mL	3.9mL	1.0mL

Table 5.9: The lowest AICc VIF corrected and reference models' cross-validation testing ME means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

Potential Models' Cross-Validated Mean Percentage Errors

Model	*Minimum*	*1st-Qt*	*Mean*	*2nd-Qt (Median)*	*3rd-Qt*	*Maximum*	*St. Dev.*
1	-10.42%	-6.61%	-1.06%	-0.99%	2.31%	8.41%	4.92%
2	-10.32%	-6.55%	-1.06%	-0.70%	2.19%	8.49%	4.88%
3	-10.48%	-6.45%	-1.07%	-0.61%	2.06%	8.41%	4.83%
4	-10.60%	-6.49%	-1.06%	-0.79%	2.19%	8.32%	4.87%
5	-10.45%	-6.53%	-1.07%	-0.65%	2.04%	8.38%	4.84%
6	-10.56%	-6.59%	-1.06%	-0.82%	2.17%	8.28%	4.88%
7	-10.87%	-6.19%	-0.94%	-0.78%	2.85%	8.57%	4.93%
8	-10.72%	-6.20%	-0.96%	-0.48%	2.34%	8.66%	4.88%
Mean	-10.55%	-6.45%	-1.04%	-0.73%	2.27%	8.44%	4.88%
St. Dev.	0.18%	0.17%	0.05%	0.15%	0.26%	0.13%	0.03%

Table 5.10: Cross-validation results suggested all 8 models have an averaged testing MPE of -1% and an averaged standard deviation of 5%. The low standard deviations between the 8 structural frameworks' testing MPE metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Mean Percentage Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	*Lowest AICc Model*	-10.42%	-6.61%	-1.06%	-0.99%	2.31%	8.41%	4.925
	Mean of All Potential Models	-10.55%	-6.45%	-1.04%	-0.73%	2.27%	8.44%	4.88%
	St. Dev. of All Potential Models	0.18%	0.17%	0.05%	0.15%	0.26%	0.13%	0.03%
All 6 Predictors Considered	*Lowest AICc Model*	-12.48%	-5.67%	-1.05%	-0.31%	2.56%	9.33%	5.09%
	Mean of All Potential Models	-12.14%	-5.60%	-1.07%	-0.41%	2.49%	9.12%	5.02%
	St. Dev. of All Potential Models	1.47%	0.80%	0.09%	0.35%	0.39%	0.65%	0.15%

Table 5.11: The lowest AICc VIF corrected and reference models' cross-validation testing MPE means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

Potential Models' Cross-Validated Mean Square Errors

Model	*Minimum*	*1st-Qt*	*Mean*	*2nd-Qt (Median)*	*3rd-Qt*	*Maximum*	*St. Dev.*
1	1317mL^2	3231mL^2	8715mL^2	10270mL^2	12280mL^2	16380mL^2	4609mL^2
2	1374mL^2	3195mL^2	8634mL^2	10280mL^2	12550mL^2	15850mL^2	4555mL^2
3	1322mL^2	3052mL^2	8646mL^2	10120mL^2	12360mL^2	16320mL^2	4601mL^2
4	1264mL^2	3086mL^2	8736mL^2	10560mL^2	12350mL^2	16900mL^2	4668mL^2
5	1336mL^2	3070mL^2	8651mL^2	10190mL^2	12330mL^2	16470mL^2	4605mL^2
6	1277mL^2	3105mL^2	8743mL^2	10630mL^2	12290mL^2	17050mL^2	4675mL^2
7	1265mL^2	3031mL^2	8781mL^2	10650mL^2	12690mL^2	16560mL^2	4685mL^2
8	1317mL^2	3001mL^2	8691mL^2	10660mL^2	12690mL^2	15990mL^2	4627mL^2
Mean	1309mL^2	3096mL^2	8700mL^2	10420mL^2	12443mL^2	16440mL^2	4628mL^2
St. Dev.	38mL^2	79mL^2	53mL^2	227mL^2	174mL^2	408mL^2	45mL^2

Table 5.12: Cross-validation results suggested all 8 models have an averaged testing MSE of 8700mL^2 and an averaged standard deviation of 4628mL^2. The low standard deviations between the 8 structural frameworks' testing MSE metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Mean Square Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	***Lowest AICc Model***	1317 mL^2	3231 mL^2	8715 mL^2	10270 mL^2	12280 mL^2	16380 mL^2	4609 mL^2
	Mean of All Potential Models	1309 mL^2	3096 mL^2	8700 mL^2	10420 mL^2	12443 mL^2	16440 mL^2	4628 mL^2
	St. Dev. of All Potential Models	38 mL^2	79 mL^2	53 mL^2	227 mL^2	174 mL^2	408 mL^2	45 mL^2
All 6 Predictors Considered	***Lowest AICc Model***	1337 mL^2	3112 mL^2	8646 mL^2	9311 mL^2	11830 mL^2	17130 mL^2	4840 mL^2
	Mean of All Potential Models	1460 mL^2	3260 mL^2	8862 mL^2	10208 mL^2	12374 mL^2	17618 mL^2	4945 mL^2
	St. Dev. of All Potential Models	247 mL^2	374 mL^2	226 mL^2	680 mL^2	635 mL^2	1727 mL^2	301 mL^2

Table 5.13: The lowest AICc VIF corrected and reference models' cross-validation testing MSE means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

Potential Models' Cross-Validated Root Mean Square Errors

Model	*Minimum*	*1st-Qt*	*Mean*	*2nd-Qt (Median)*	*3rd-Qt*	*Maximum*	*St. Dev.*
1	36.3mL	56.8mL	89.1mL	101.3mL	110.8mL	128.0mL	27.9mL
2	37.1mL	56.5mL	88.7mL	101.4mL	112.0mL	125.9mL	27.8mL
3	36.4mL	55.2mL	88.7mL	100.6mL	111.2mL	127.8mL	28.0mL
4	35.6mL	55.6mL	89.2mL	102.7mL	111.1mL	130.0mL	28.1mL
5	36.6mL	55.4mL	88.7mL	101.0mL	111.0mL	128.3mL	28.0mL
6	35.7mL	55.7mL	89.2mL	103.1mL	110.8mL	130.6mL	28.1mL
7	35.6mL	55.1mL	89.3mL	103.2mL	112.6mL	128.7mL	28.3mL
8	36.3mL	54.8mL	88.9mL	103.2mL	112.7mL	126.5mL	28.2mL
Mean	36.2mL	55.6mL	89.0mL	102.1mL	111.5mL	128.2mL	28.0mL
St. Dev.	0.5mL	0.7mL	0.3mL	1.1mL	0.8mL	1.6mL	0.2mL

Table 5.14: Cross-validation results suggested all 8 models have an averaged testing RMSE of 89mL and an averaged standard deviation of 28mL. The low standard deviations between the 8 structural frameworks' testing RMSE metric means, medians, and individual model standard deviations suggested the models were of similar performance and robustness. There were no strong indicates that any of the 8 structural frameworks were either far superior or inferior in allograft TCV prediction.

Root Mean Square Errors: Summary and Comparison

	Model	Minimum	1st-Qt	Mean	2nd-Qt (Median)	3rd-Qt	Maximum	St. Dev.
Predictors Considered by VIF	*Lowest AICc Model*	36.3mL	56.8mL	89.1mL	101.3mL	110.8mL	128.0mL	27.9mL
	Mean of All Potential Models	36.2mL	55.6mL	89.0mL	102.1mL	111.5mL	128.2mL	28.0mL
	St. Dev. of All Potential Models	0.5mL	0.7mL	0.3mL	1.1mL	0 .8mL	1.6mL	0.2mL
All 6 Predictors Considered	*Lowest AICc Model*	36.6mL	55.8mL	88.5mL	96.5mL	108.7mL	130.9mL	28.7mL
	Mean of All Potential Models	38.1mL	57.0mL	89.6mL	101.0mL	111.2mL	132.6mL	29.0mL
	St. Dev. of All Potential Models	3.2mL	3.2mL	1.2mL	3.4mL	2.9mL	6.4mL	1.0mL

Table 5.15: The lowest AICc VIF corrected and reference models' cross-validation testing MAE means, medians, and standard deviations might have suggested the reference modeling process produced the "best" prediction model. However, the improved performance of the reference modeling process, with respect to the VIF correction process, was generally minimal, was not consistent for all models, and came at the cost of increased error variance.

The 8 top potential models were investigated to determine if the specifications needed to be refined to account for heteroscedasticity. The Breusch-Pagan test results rejected the null hypothesis of homoscedasticity, i.e., heteroscedasticity was suggested, for all 8 models (p-values << 0.05) and the results were presented in Table 5.16. The reported Table 5.16 results were determined using the REML to estimate the coefficients and therefore the reported AICc values were neither comparable to the earlier "glmulti" AICc values nor between the models within Table

5.16 due to the varied structural frameworks. However, the AICc values were presented in Tables 5.16 and 5.17 to allow for a simple investigate in determining if heteroscedasticity corrections improved the overall prediction performances of a specific structural framework.

Testing for Heteroscedasticity

Model	N	Breusch-Pagan test P-Value	AICc
1	97	0.01	-63.82
2	97	0.01	-60.56
3	97	0.01	-59.49
4	97	0.01	-62.71
5	97	0.01	-62.47
6	97	0.01	-65.69
7	97	0.01	-61.98
8	97	0.01	-58.74

Table 5.16: The Breusch-Pagan test suggested strong heteroscedasticity for all 8 models (p-values << .05) that very likely needed to be addressed to improve model performance. The presented p-values were not exactly equal to 0.01 but rather this reflects a round off error in the precision. The results were determined using the REML.

Two methods to correct for heteroscedasticity were considered in step 5. The Breusch-Pagan test and AICc value results, reflecting after the heteroscedasticity corrections were implemented, were included in Tables 5.17.

The first correction method used the *R* “varClass” functions to correct heteroscedasticity by applying weights to re-estimate the coefficients. The coefficients were re-estimated such that the models were made more robust to heteroscedasticity. The “varClass” functions results in Table 5.17 did generally show negligible p-value improvements for the Breusch-Pagan test, howver, they were outside the reported precision in Tables 5.16 and 5.17. The slight decrease in the AICc values between the same structural frameworks in Tables 5.16 and 5.17 did support the implementation

of the “varClass” functions did improve the overall models’ performances. Although improvements were observed, the Breusch-Pagan test still detected heteroscedasticity in all 8 models.

The second correction method used a Cook’s distance correction method to identify and remove influential outliers that might have caused the measurable heteroscedasticity presented in the models. The Cook’s distance method results in Table 5.17 showed large increases in the Breusch-Pagan test p-values and large decreases in the AICc values. These large changes in the p-values and AICc values strongly suggested the Cook’s distance method was a much more effective method at addressing the heteroscedasticity in the data and the models’ overall performances. In fact, Cook’s distance method performed so well at addressing heteroscedasticity that the Breusch-Pagan test failed to reject the null hypothesis of homoscedasticity for all 8 models. These superior, intermediate results justified the use of the Cook’s distance method to correct the heteroscedasticity issues for all 8 top models.

Heteroscedasticity Correction Results

	"varClass" Weight Corrected			Cook's Distance Corrected		
Model	N	Breusch-Pagan test P-value	AICc	N	Breusch-Pagan test P-value	AICc
1	97	0.01	-66.16	92	0.73	-85.08
2	97	0.01	-62.96	92	0.69	-83.14
3	97	0.01	-61.86	92	0.71	-82.32
4	97	0.01	-64.91	92	0.74	-84.13
5	97	0.01	-64.69	92	0.72	-85.42
6	97	0.01	-68.06	92	0.75	-87.25
7	97	0.01	-64.51	91	0.49	-86.13
8	97	0.01	-61.43	91	0.49	-84.40

Table 5.17: Breusch-Pagan test and AICc values demonstrated, when compared to the Table 5.16 results, the "varClass" and Cook's distance methods helped to correct the heteroscedasticity and improve the overall models' performances. The "varClass" Breusch-Pagan test p-values did generally improve but the quantitative change was outside the reported precision. The Cook's distance method performed superiorly in correcting the heteroscedasticity issues and improving overall model's performances. The results were determined using the REML.

After Cook's distance was shown to improve heteroscedasticity detected in the data, the 8 top potential models had their coefficients re-estimated using REML and Cook's distance. In addition to addressing heteroscedasticity in the data, the Cook's distance method also helped to remove influential outliers that could negatively impact the final model. The 8 top potential model structural frameworks and coefficient estimates were presented in Tables 5.18 to 5.25. As Table 5.17 indicates, these 8 models estimated coefficients were trained with either 92 or 91 data points, i.e., Cook's distance identified 5 or 6 influential outliers to remove in the modeling process. Once the models' coefficients were re-estimated, the coefficients were investigated to determine if they were statistically-significant in step 6.

All estimated coefficients for Models 1 to 6, i.e., Tables 5.18 to 5.23, were statistically-significant (p-values << 0.025). There were estimated coefficients in Models 7 and 8, i.e., Tables 5.24 and 5.25, that were statistically non-significant (p-values >> 0.025) and therefore could

negatively affect the models' robustness. The non-significant estimated coefficients in Models 7 and 8 were terms containing the Sex variable. The larger reported estimated coefficient SEs and smaller t-value magnitudes for Models 7 and 8 supported these models were less robust than for Models 1 to 6 and therefore justified the insignificant terms should be considered for removal.

Iteratively removing the largest p-values ≥ 0.025 demonstrated Models 7 and 8 were permutations of Models 4 and 3, respectively. As was expected, removing the non-significant estimated coefficients improved the remaining estimated coefficient SEs and t-value magnitudes of Model 7 and 8 because their structural frameworks became that of Models 4 and 3, respectively. The *ln(Ht)* estimated coefficients, SEs, and t-value magnitudes for the 8 top VIF corrected models were compiled in Table 5.27 (the table will be further discussed shortly) to help investigate the robustness of the models. The larger *ln(Ht)* estimated coefficient SEs and smaller t-value magnitudes for Models 7 and 8 further demonstrated that these models were less robust than Models 1 to 6. As was expected, the SEs and t-value magnitudes for *ln(Ht)* in Models 7 and 8 were seen to improve in Table 5.27 when their structural frameworks reduced to Models 4 and 3, respectively. Given the removal of the insignificant coefficient terms reduced the structural frameworks of Models 7 and 8 to Models 4 and 3, respectively, there were only 6 final models of interest and no need to re-estimate the coefficients a final time. Although the results suggested Models 7 and 8 were no longer models of interest to predict allograft TCV, they continued to be analyzed herein to confirm the intermediate finding.

The lowest "glmulti" AICc reference model's structural framework and estimate coefficients was presented in Table 5.26 such that it could be compared to Tables 5.18 to 5.25. Given the lowest AICc reference model had large estimate coefficient SEs, low t-value magnitudes, and statistically-insignificant p-values (> 0.05) suggest the reference model was not robust. For a closer comparison of model robustness between the 8 top VIF and the 8 top non-VIF corrected models, Table 5.27 (briefly mentioned) was created in which the *ln(Ht)* estimated coefficients, SEs, and t-value magnitudes were isolated because they were present in all 16 models. The table shows the *ln(Ht)* estimated coefficients, SEs and t-value magnitudes for the 8 top models were relatively stable and did not show signs of classical multicollinearity issues. However, the *ln(Ht)* estimated

coefficients, SEs, and t-value magnitudes for the reference set were unstable – this was a classical sign of strong multicollinearity [108,109]. This finding appeared to confirm the earlier indication that failing to account for multicollinearity in the reference modeling process reduced these models' robustness.

Before moving on to identify the final, "best" allograft TCV prediction model in this initial modeling process, instructions on how to use these models are warranted. To predict TCV from any of these 9 models, i.e., Tables 5.18 to 5.26, the user is to use the appropriate "Modeled *log-log* Transform Equation". The inputted units for Age, Ht, Wt, BSA, and BMI are required to be months, cm, kg, m^2, and kg/m^2, respectively. Sex is a dummy variable in which male is numerically code as "1" and female is "0". The model outputs the predicted allograft TCV in mL. The *ln( )* and *exp( )* functions forward transform the predictors and backward transform the TCV, respectively.

Final Results for Initial Modeling Process: Model 1

Structural Framework (R software notation)				
TCV ~ 1 + Ht + BMI + BMI:Sex				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) + \boldsymbol{\alpha_4} * \ln(\text{BMI}) * \text{Sex}$ ***or*** $\text{TCV} = \exp(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) + \boldsymbol{\alpha_4} * \ln(\text{BMI}) * \text{Sex})$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-5.509098	0.26960075	-20.43428	0.00E+00
$\boldsymbol{\alpha_2}$	2.072416	0.06270071	33.05252	0.00E+00
$\boldsymbol{\alpha_3}$	0.41489	0.05708467	7.26798	0.00E+00
$\boldsymbol{\alpha_4}$	0.038364	0.00914381	4.19566	1.00E-04

Table 5.18: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (<< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht and BMI increases with TCV. The interaction term of BMI with Sex suggested BMI had an additional, positive effect on the TCVs of males.

Final Results for Initial Modeling Process: Model 2

Structural Framework (R software notation)				
TCV ~ 1 + Ht + BMI:Sex + Ht:BMI				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-4.325414	0.30333033	-14.259746	0.00E+00
$\boldsymbol{\alpha_2}$	1.834874	0.08288294	22.138135	0.00E+00
$\boldsymbol{\alpha_3}$	0.037606	0.00908287	4.140354	1.00E-04
$\boldsymbol{\alpha_4}$	0.083182	0.01122548	7.410076	0.00E+00

Table 5.19: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (<< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht increases with TCV. The interaction term of BMI with Sex suggested BMI had an additional, positive effect on the TCVs of males. The interaction term of BMI with Ht suggested BMI had a larger, positive effect on the TCVs of taller individuals.

Final Results for Initial Modeling Process: Model 3

Structural Framework (R software notation)				
TCV ~ 1 + Ht + Ht:Sex + Ht:BMI				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{Ht}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{Ht}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-4.25571	0.30258291	-14.064608	0.00E+00
$\boldsymbol{\alpha_2}$	1.806543	0.08290397	21.790791	0.00E+00
$\boldsymbol{\alpha_3}$	0.023339	0.0056087	4.161122	1.00E-04
$\boldsymbol{\alpha_4}$	0.087798	0.01119582	7.841997	0.00E+00

Table 5.20: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (<< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht increases with TCV. The interaction term of Ht with Sex suggested Ht had an additional, positive effect on the TCVs of males. The interaction term of BMI with Ht suggested BMI had a larger, positive effect on the TCVs of taller individuals.

Final Results for Initial Modeling Process: Model 4

Structural Framework (R software notation)				
TCV ~ 1 + Ht + BMI + Ht:Sex				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \text{Sex}$ ***or*** $\text{TCV} = \exp(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{BMI}) + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \text{Sex})$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-5.504916	0.26956228	-20.42169	0.00E+00
$\boldsymbol{\alpha_2}$	2.056973	0.0628142	32.74695	0.00E+00
$\boldsymbol{\alpha_3}$	0.438429	0.05703488	7.68704	0.00E+00
$\boldsymbol{\alpha_4}$	0.02374	0.00565636	4.19708	1.00E-04

Table 5.21: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (<< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht and BMI increases with TCV. The interaction term of Ht with Sex suggested Ht has an additional, positive effect on the TCVs of males.

Final Results for Initial Modeling Process: Model 5

Structural Framework (R software notation)				
TCV ~ 1 + Sex + Ht + Ht:BMI				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \ln(\text{BMI})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-4.352287	0.30370611	-14.330586	0.00E+00
$\boldsymbol{\alpha_2}$	0.115883	0.02795277	4.145687	1.00E-04
$\boldsymbol{\alpha_3}$	1.826155	0.08283307	22.046205	0.00E+00
$\boldsymbol{\alpha_4}$	0.087737	0.01120223	7.832095	0.00E+00

Table 5.22: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (<< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht increases with TCV. The model also suggested TCVs were larger in males. The interaction term of BMI with Ht suggested BMI had a larger, positive effect on the TCVs of taller individuals.

Final Results for Initial Modeling Process: Model 6

Structural Framework (R software notation)				
TCV ~ 1 + Sex + Ht + BMI				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{BMI})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{BMI})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-5.602321	0.27121305	-20.65653	0.00E+00
$\boldsymbol{\alpha_2}$	0.117915	0.0281886	4.18306	1.00E-04
$\boldsymbol{\alpha_3}$	2.076736	0.06273797	33.10174	0.00E+00
$\boldsymbol{\alpha_4}$	0.438145	0.05706414	7.67811	0.00E+00

Table 5.23: The model's final estimated coefficients were trained with N = 92 healthy heart data points. All estimated coefficient p-values were statistically-significant (< 0.025) and suggested the model estimates were robust. The model suggested an individual's Ht and BMI increases with TCV. The model also suggested TCVs were larger in males.

Final Results for Initial Modeling Process: Model 7

Structural Framework (R software notation)				
TCV ~ 1 + Sex + Ht + BMI + Sex:Ht				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{BMI}) + \boldsymbol{\alpha_5} * \text{Sex} * \ln(\text{Ht})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \text{Sex} + \boldsymbol{\alpha_3} * \ln(\text{Ht}) + \boldsymbol{\alpha_4} * \ln(\text{BMI}) + \boldsymbol{\alpha_5} * \text{Sex} * \ln(\text{Ht})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-5.138692	0.6436249	-7.983985	0
$\boldsymbol{\alpha_2}$	-0.387835	0.7059376	-0.54939	0.5842
$\boldsymbol{\alpha_3}$	1.993095	0.1330423	14.980915	0
$\boldsymbol{\alpha_4}$	0.4258	0.0562411	7.570971	0
$\boldsymbol{\alpha_5}$	0.09972	0.14162	0.704141	0.4832

Table 5.24: The model's final estimated coefficients were trained with N = 91 healthy heart data points. Two of the estimated coefficient p-values were not statistically-significant (> 0.05) and suggested not all model estimates were robust. Setting the coefficient with the largest p-value, i.e., $\boldsymbol{\alpha_2}$, to zero effectively removed that corresponding term from the model. Removing the $\boldsymbol{\alpha_2}$ related term converted Model 7 to Model 4. This result suggested Model 7 was a less robust permutation of Model 4. The model had insignificant terms and therefore the coefficients were not interpreted.

Final Results for Initial Modeling Process: Model 8

Structural Framework (R software notation)				
TCV ~ 1 + Sex + Ht + Sex:Ht + Ht:BMI				
Modeled log-log Transform Equation				
$\ln(\mathrm{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \mathrm{Sex} + \boldsymbol{\alpha_3} * \ln(\mathrm{Ht}) + \boldsymbol{\alpha_4} * \mathrm{Sex} * \ln(\mathrm{Ht}) + \boldsymbol{\alpha_5} * \ln(\mathrm{Ht}) * \ln(\mathrm{BMI})$ ***or*** $\mathrm{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \mathrm{Sex} + \boldsymbol{\alpha_3} * \ln(\mathrm{Ht}) + \boldsymbol{\alpha_4} * \mathrm{Sex} * \ln(\mathrm{Ht}) + \boldsymbol{\alpha_5} * \ln(\mathrm{Ht}) * \ln(\mathrm{BMI})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	-3.906945	0.6547818	-5.966788	0
$\boldsymbol{\alpha_2}$	-0.408828	0.6997865	-0.584219	0.5606
$\boldsymbol{\alpha_3}$	1.745897	0.1428938	12.218146	0
$\boldsymbol{\alpha_4}$	0.103541	0.1403862	0.737544	0.4628
$\boldsymbol{\alpha_5}$	0.085355	0.0110317	7.737252	0

Table 5.25: The model's final estimated coefficients were trained with N = 91 healthy heart data points. Two of the estimated coefficient p-values were not statistically-significant (> 0.05) and suggested not all model estimates were robust. Setting the coefficient with the largest p-value, i.e., $\boldsymbol{\alpha_2}$, to zero effectively removed that corresponding term from the model. Removing the $\boldsymbol{\alpha_2}$ related term converted Model 8 to Model 3. This result suggested Model 8 was a less robust permutation of Model 3. The model had insignificant terms and therefore the coefficients were not interpreted.

Final Results for Reference Model with Lowest AICc Rank

Structural Framework (R software notation)				
TCV ~ 1 + Ht + Age:Sex + Ht:Sex + Ht:Wt				
Modeled log-log Transform Equation				
$\ln(\text{TCV}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{Age}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \text{Sex} + \boldsymbol{\alpha_5} * \ln(\text{Ht}) * \ln(\text{Wt})$ ***or*** $\text{TCV} = \exp\left(\boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} * \ln(\text{Ht}) + \boldsymbol{\alpha_3} * \ln(\text{Age}) * \text{Sex} + \boldsymbol{\alpha_4} * \ln(\text{Ht}) * \text{Sex} + \boldsymbol{\alpha_5} * \ln(\text{Ht}) * \ln(\text{Wt})\right)$				
Coefficients	***Value***	***SE***	***t-value***	***p-value***
$\boldsymbol{\alpha_1}$	0.2576417	0.7809469	0.329909	0.7423
$\boldsymbol{\alpha_2}$	0.8502569	0.192942	4.4068	0
$\boldsymbol{\alpha_3}$	0.0771641	0.0399931	1.929435	0.057
$\boldsymbol{\alpha_4}$	-0.0549694	0.0389858	-1.409984	0.1622
$\boldsymbol{\alpha_5}$	0.0840404	0.0107574	7.812332	0

Table 5.26: The model's final estimated coefficients were trained with N = 91 healthy heart data points. Three of the estimated coefficient p-values were not statistically-significant (> 0.05) and suggested the model estimates were not robust. The fact that this model (1) had the lowest AICc value, (2) had statistically-insignificant estimate coefficients, and (3) did not undergo VIF correction, suggested that failing to account for multicollinearity might have affected the robustness of the models derived in the reference modeling procedure. The model had insignificant terms and therefore the coefficients were not interpreted.

Final ln(Ht) Coefficient Comparison for the 8 Top AICc Models

	VIF Corrected			***Reference (non-VIF Corrected)***		
Model	***Coefficient Value***	***SE***	***t-value***	***Coefficient Value***	***SE***	***t-value***
1	2.072	0.063	33.053	0.850	0.193	4.407
2	1.835	0.083	22.138	0.740	0.203	3.646
3	1.807	0.083	21.791	1.388	0.422	3.289
4	2.057	0.063	32.747	1.262	0.422	2.991
5	1.826	0.083	22.046	2.635	0.324	8.135
6	2.077	0.063	33.102	0.844	0.194	4.342
7	1.993	0.133	14.981	0.760	0.234	3.247
8	1.746	0.143	12.218	1. 338	0. 346	3.868
Mean	1.927	0.089	24.010	1.211	0.285	4.241
St. Dev.	0.137	0.032	8.226	0.677	0.104	1.654

Table 5.27: The 8 lowest AICc value models for both the VIF corrected and reference modeling procedures were presented with their estimated coefficients, SEs, and t-value magnitudes. A one-way ANOVA confirmed the *ln(Ht)* estimated coefficients (p-value = 0.0082, effect size = 1.46) and SEs (p-value ≤ 0.0001, effect size = 2.54) were statistically different between the modeling procedures. The statistical difference between the two modeling procedure estimate coefficients and the reference procedure's larger SEs and smaller t-value magnitudes suggested there is a large, classical multicollinearity issue in the reference models. Finding that the reference models suffer from multicollinearity was expected given the VIF procedure was intentionally skipped.

Hereto multicollinearity, heteroscedasticity, and influential outlier issues were tested and corrected for in the modeling process. The 8 models' coefficients were already re-estimated using the REML method with all 97 data points (excluding the influential outliers identified by Cook's distance) and presented in Tables 5.18 to 5.25. The 8 models' final testing statistical modeling metrics were re-estimated with a second, final cross-validation in which the modeling corrections were included. The final testing metric results were presented in Table 5.28. The need to cross-validate the final testing statistical model metric results derived from the limited data available to formally set aside for testing with the recognition that training errors do not represent testing errors. The final analysis of the testing statistical modeling metric results was needed to (1) determine if

the decreased robustness of Models 7 and 8 had an effect on the Models' predictive capabilities and (2) identify a final model for predicting allograft TCVs from donor parameters. It is worth explicitly stating that the second, final cross-validation procedure did not use the estimate coefficients in Tables 5.18 to 5.25 to estimate the final testing statistical model metric results.

The initial inspection of the 48-individual testing statistical model metric results in Table 5.28 demonstrated a challenge in identifying the "best" model. The challenge in identifying the "best" model was there were minor differences in the models' errors and the models' ranking order fluctuated, based on which specific metric was being analyzed. Therefore, the results in Table 5.28 were reported in Table 5.29 as a numeric ranking based on the sequential order for each model within a given metric. A value of 1 was given to the model with the most preferable metric value for that given metric. For all statistical modeling metrics, values approaching zero were preferable. Table 5.29 presents the final, model sequential ranks for each metric (numeric, color-coded) and the overall model performance (row, color-coded). Red, green, and blue indicate 1st, 2nd, and 3rd place rankings. To simplify the model ranking assessments, the ranking means, medians, and totals for each of the statistical models were reported such that each metric was given an equal weighting of consideration.

Models 2 and 5 dominated 1st and 2nd places in Table 5.29. Models 3 and 8 fluctuated between 3rd and 4th place, this was an interesting find because Model 8 is a permutation of Model 3. Model 8 had a lower mean and total ranking while Model 3 had a lower "glmulti" AICc and better median, MSE, and RMSE rankings. In fact, Model 3 outperformed Model 8 in ranking for all test statistical metrics except for the testing ME and MPE results. The MSE and RMSE are a special set of metrics are special set metrics are often used or recommended for use to analysis model errors because they increase the penalty for larger errors [108,118,133]. Furthermore, the MAPE and MPE are asymmetric error metrics that have an under-prediction penalty bias (discussed later). To address the under-prediction penalty bias, Table 5.30 presents the symmetrically corrected MAPE and MPE values, i.e. sMAPE and sMPE values. Although Model 3 and 8's MAPE ranking showed no major change, there was a large change between the MPE and sMPE rankings. In

replacing the MAPE and MPE values for the sMAPE and sMPE values, the ranking means, medians, and totals strengthened Models' 2, 5, and 3's rankings. Given:

1) Model 8 was a permutation of Model 3 with insignificant coefficients,
2) Model 8 had larger testing metric error variances (as demonstrated in step 4),
3) Model 3's ranking was greatly reduced when MPE was replaced with sMPE, and
4) Model 3 outperformed in MSE and RMSE rankings,

it was concluded that Model 3 outperformed Model 8. Therefore, Model 3 took a ranking of 3rd place in the modeling process. To appreciate the final model selection process, an overview of what error attributes were being highlighted by each of the metrics will be discussed in the next section.

Model 2 was ranked 1st in the selection process and therefore it was ultimately determined to be the final, i.e., "best", allograft TCV prediction model derived from the initial modeling process. Model 2 was interchangeable referred to as Model A because it was the final model in the initial modeling process. The Model 2 terminology was generally reserved henceforth when comparisons of Model A to the other 7 initial, potential models were made, otherwise the Model A terminology was preferential. The term "A" was used because after a preliminary assessment of the initial modeling procedural results a few modeling procedural changes were made to try and improve the final allograft TCV prediction model in an upcoming section.

Final Testing Statistical Model Metric Results

Model	*MAE*	*MAPE*	*ME*	*MPE*	*MSE*	*RMSE*
1	64.220mL	12.253%	5.801mL	-1.037%	8578mL^2	92.620mL
2	63.795mL	12.190%	5.431mL	-1.079%	8497mL^2	92.180mL
3	63.956mL	12.211%	5.732mL	-1.088%	8552mL^2	92.478mL
4	64.411mL	12.279%	6.112mL	-1.035%	8646mL^2	92.983mL
5	63.870mL	12.187%	5.643mL	-1.075%	8536mL^2	92.389mL
6	64.416mL	12.275%	6.065mL	-1.036%	8647mL^2	92.991mL
7	64.799mL	12.487%	5.894mL	-0.956%	8694mL^2	93.243mL
8	64.107mL	12.356%	5.484mL	-1.012%	8566mL^2	92.552mL

Table 5.28: The presented results were the final testing statistical model metric results used to identify the final prediction model in the initial modeling protocol. Starting with N = 97 data points, results were determined with 10-fold cross-validation, REML, and Cook's distance.

Final Ranked Testing Statistical Model Metric Results

Model	*MAE*	*MAPE*	*ME*	*MPE*	*MSE*	*RMSE*	*Mean*	*Median*	*Total*
1	5	4	5	5	5	5	4.8	5.0	29.0
2	**1**	**2**	**1**	7	**1**	**1**	**2.2**	**1.0**	**13.0**
3	**3**	**3**	4	8	**3**	**3**	4.0	**3.0**	24.0
4	6	6	8	**3**	6	6	5.8	6.0	35.0
5	**2**	**1**	**3**	6	**2**	**2**	**2.7**	**2.0**	**16.0**
6	7	5	7	4	7	7	6.2	7.0	37.0
7	8	8	6	**1**	8	8	6.5	8.0	39.0
8	4	7	**2**	**2**	4	4	**3.8**	4.0	**23.0**

Table 5.29: To ease interpretation of the final testing statistical model metric results, models were given a ranking of 1 to 8 for each individual metric in which 1 was considered "best". Colored numbers and rows indicated 1st (red), 2nd (green), and 3rd (blue) placed models in performance by specific metric and overall model, respectively. Metric ranking means, medians, and totals were taken to summarize model performance rankings. Model 2 generally outperformed all other models; this finding suggests Model 2 was the "best" allograft TCV prediction model within the initial modeling process. Model 2's MSE and RMSE were ranked 1stl, this result further supports Model 2 was the "best" model.

Final Ranked Testing Statistical Model Metric Results with sMAPE and sMPE

Model	sMAPE	sMAPE	MAPE	sMPE	sMPE	MPE	Mean	Median	Total
	Value	Rank	Rank	Value	Rank	Rank	Rank	Rank	Rank
1	12.08%	4	4	0.16%	4	5	4.7	5.0	28.0
2	12.01%	1	2	0.12%	2	7	1.2	1.0	7.0
3	12.03%	3	3	0.11%	1	8	2.8	3.0	17.0
4	12.11%	5	6	0.16%	5	3	6.0	6.0	36.0
5	12.02%	2	1	0.12%	3	6	2.3	2.0	14.0
6	12.11%	6	5	0.16%	6	4	6.7	7.0	40.0
7	12.35%	8	8	0.29%	8	1	7.7	8.0	46.0
8	12.21%	7	7	0.22%	7	2	4.7	4.0	28.0

Table 5.30: The asymmetrical MAPE and MPE results were replaced in the current table with the sMAPE and sMPE metrics. The sMAPE and sMPE metrics address a well-established under-prediction penalty bias in the MAPE and MPE metrics. The ranking means, medians, and totals were recalculated after the MAPE and MPE metrics were replaced. The testing statistical model metric symmetry corrections increased support for Models 2, 5, and 3 to be ranked 1st, 2nd, and 3rd, respectively. Interestingly, by making the symmetric testing statistical model metric corrections, the MPE related metric rankings for Models 2, 5, and 3 (i.e., 1st, 2nd, and 3rd place models) went from 7th, 6th, and 8th place to 2nd, 3rd, and 1st place.

Graphical representations of the final 10-fold cross-validated prediction results for Model A were presented with respect to the mTCVs in Figures 5.3 and 5.4. The pTCV values were determined by averaging the 100 10-fold cross-validation prediction results. Visual inspection of these figures generally suggested the errors fluctuate about the slope (line) of unity, i.e., the solid, black line. Model A's residual distribution was presented in Figures 5.5 and 5.6 with both a quantile-quantile plot and histogram plot, respectively. Although visual inspection of the histogram in Figure 5.6 might be interrupted as being normal distributed; the normal quantile-quantile map presents "tails" suggested the residuals were not normality distributed. The null hypothesis of normality was rejected (p-value = 0.0002) with a Shapiro-Wilk test and therefore generally confirmed the errors were not normally distributed. The visual inspection of these residual plots suggested only mild deviation from normality, as is typically expected in real world empirical data. There was no

perceived indication this discrepancy from the normality assumption harmed the overall prediction capability of Model A and the overall initial modeling process.

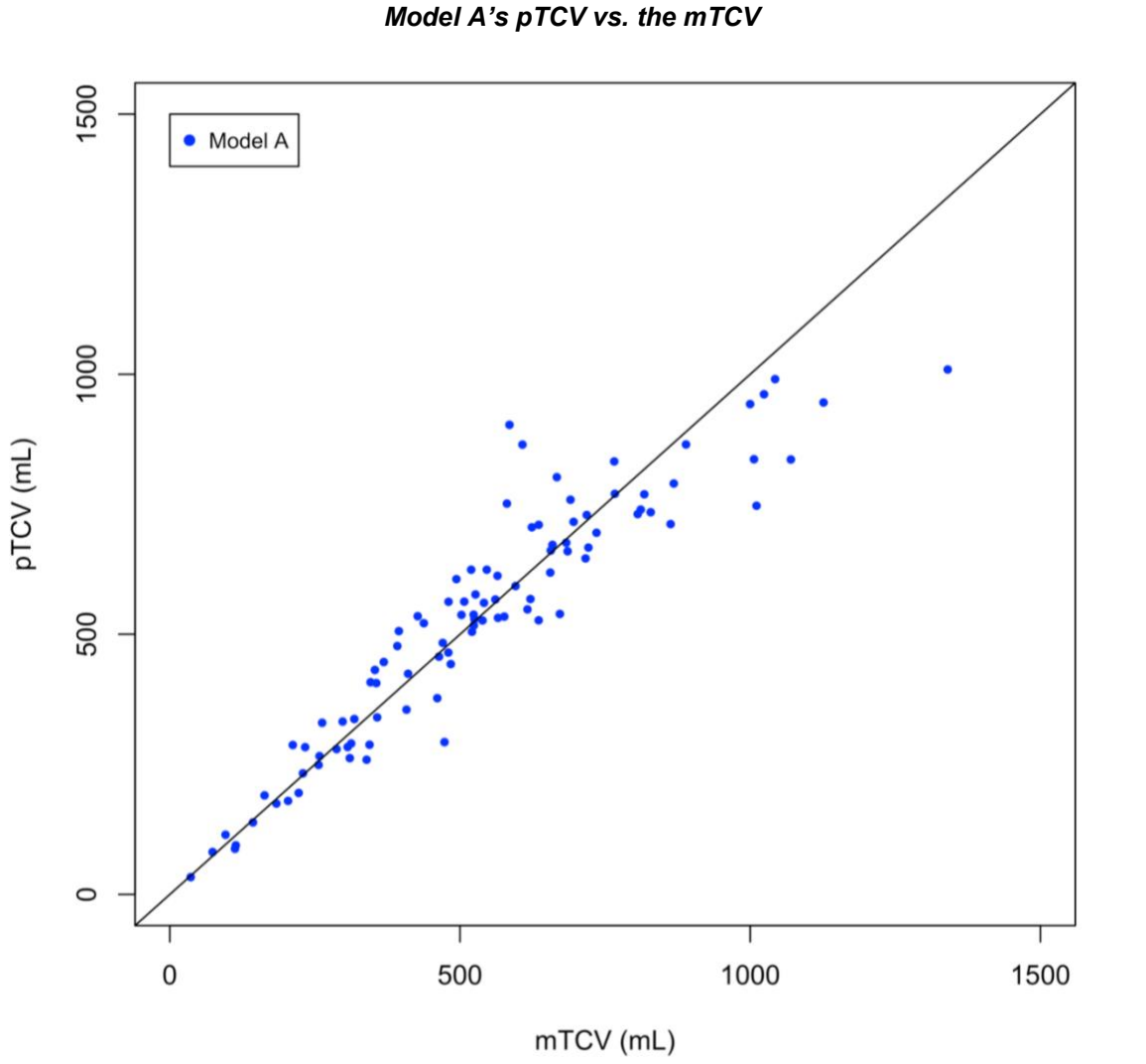

Figure 5.3: The 10-fold cross-validated pTCV values (N = 97) for Model A were plotted against their corresponding mTCV values in a "predicted vs. actual" graph. Allograft TCV data points above or below the diagonal line indicated over- or under-predictions, respectively. Visual inspection of the graph suggested (1) Model A's absolute error increased as the TCV increased and (2) Model A has a specific bias for under predicting in the largest of TCVs.

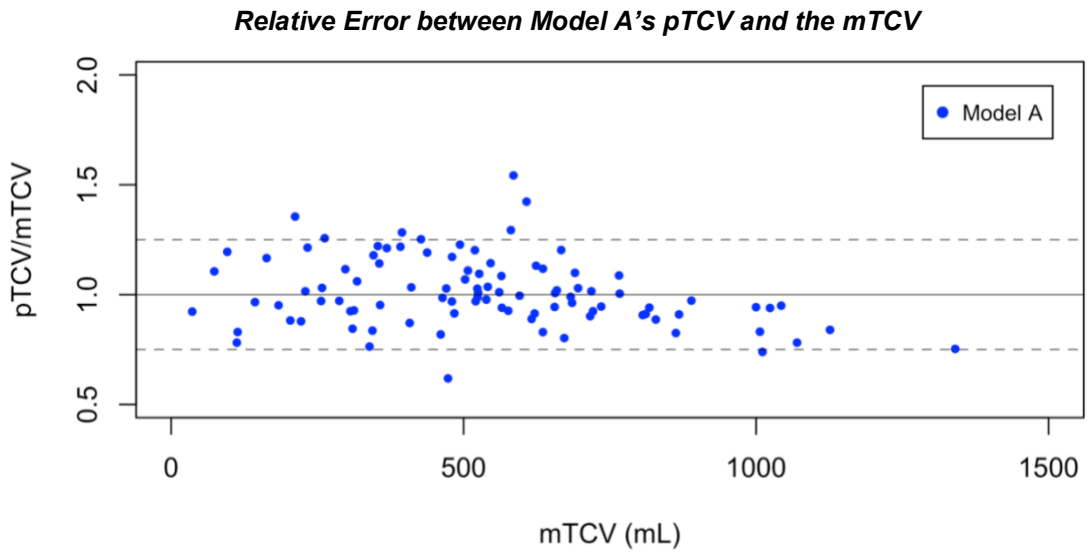

Figure 5.4: Model A pTCV values (N = 97) were represented in a "relative error" graph to visualize how the prediction errors propagated as the allograft mTCVs increased. Visual inspection suggested no relative error bias until the mTCVs had reached a volume of approximately 750mL or greater – was also seen in Figure 5.3. Although the errors of Model A increased, as was shown in Figure 5.3, visual inspection of the current Figure suggested the relative errors held relatively constant. Dashed lines show ±25% relative error bandwidth.

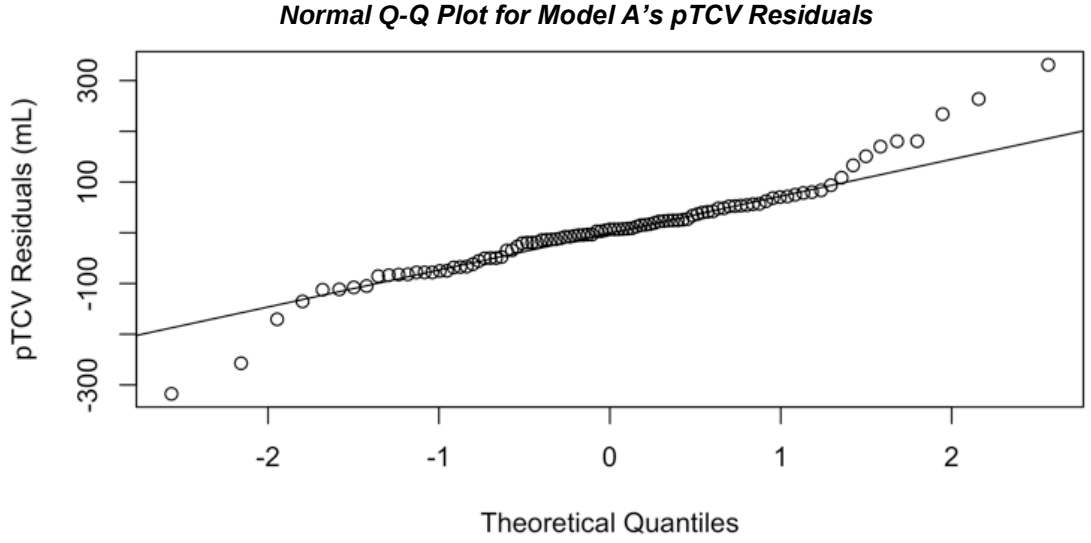

Figure 5.5: A normal Quantiles-Quantiles plot was presented to investigate the normality assumption of the Model A pTCV residuals. These plots suggest the increased likelihood of normality as the residuals better fit a single, linear line. The residuals herein generally fitted a linear line with the exception of visually appearing mild "tails" to suggested a near-normal distribution was present. A Shapiro-Wilk test rejected the null hypothesis of normality (p-value = 0.0002) to further confirm the lack of normality.

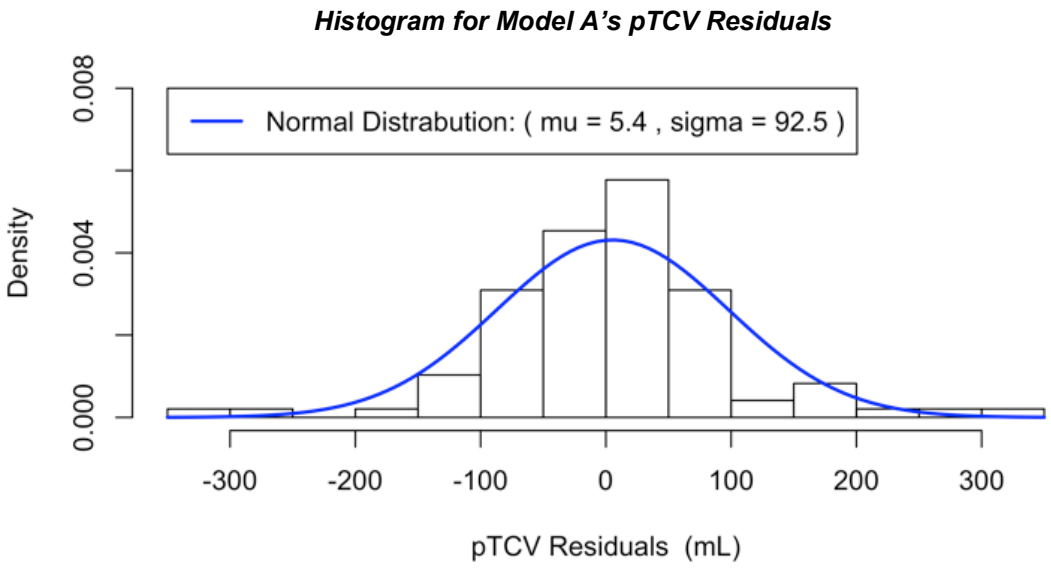

Figure 5.6: A frequency plot was created with the Model A pTCV residuals placed in one of 14 histogram containers. Visual inspection of the plot suggested either the residuals had a normal or near-normal distribution. The continuous curve (blue) was included to present an actual normal distribution with a mean (5.4mL) and standard deviation (92.5mL) matching the residual data. A Shapiro-Wilk test rejected the null hypothesis of normality (p-value = 0.0002).

5.3 Discussion of the Allograft TCV prediction Model

Important topics related to the development and validation of the initial allograft TCV prediction model, i.e., Model A, were discussed in the current section. The topics of interest included the exhaustive search, cross-validation, multicollinearity, heteroscedasticity, metric analysis, and the overall model development process. In later sections of the current chapter, the development and validation of Model B were covered and compared to Model A.

Model A's modeling process used the "glmulti" function to implement a so called "exhaustive search". The "glmulti" function searched for possible allograft TCV prediction model structural frameworks to be considered during model develop and had their corresponding AICc values calculated. A true exhaustive search would have considered all structural frameworks, however, the "glmulti" function's algorithm included 2 simplifications that made the process a "pseudo" exhaustive search. The simplifications were likely included to cut down on the algorithm's computational cost - a limitation that was highlighted in the function's documentation [122].

First, the algorithm only considered models with an intercept in the structural framework. The algorithm did not consider models with a forced non-intercept condition, i.e., a condition that automatically set the intercept's coefficient to zero. Although the algorithm did not prevent the function from estimating the intercept's coefficient to be exactly zero, nuisance factors in the real-world made this an unlikely scenario. Near-zero intercept estimates could have set to zero by the author but this post-processing procedure would have been a subjective judgement call.

Second, the algorithm only considered structural frameworks with no greater than pairwise, i.e., 2-way, interactions terms. Pairwise and higher interaction terms could have been pre-processed in theory, e.g., $x_1 x_2 = x_{12}$, and then included as additional "main effect" terms, e.g., x_{12}. Including these pre-processed, interaction terms as "main effects" would have allowed the function to consider higher-level interactions in the modeling process. However, there were several issues in how the proposed method would have forced the higher-level interactions: (1) higher-order predictor powers would have been included, e.g., $x_{12} x_1 = x_1 x_1 x_2 = {x_1}^2 x_2$, (2) computational cost would have increased, and (3) "spurious" multicollinearity could have been introduced into the model. The "spurious" multicollinearity refers to higher-level interaction terms having strong linear relationships with other terms – it is "spurious" because the main effects, having undergone the VIF procedure, did not have linear relationships until the higher interactions were included [134]. Furthermore, introducing higher-order predictor powers would, at a minimum, have complicated the field of allometry concepts.

The discussion on the "glmulti" function has demonstrated the function implemented in this body of work was not a true exhaustive search algorithm. The simplifications helped to reduce computational cost and resulted in the prevention of an end-user developing "spurious" multicollinearity in their modeling process. Even with the two well-needed simplifications included, the "glmulti" function, for all intents and purposes, was an exhaustive search – especially when compared to the "forward", "backward", and "stepwise" model selection methodologies. Given a careful consideration of these facts, the author decided to continue referring to the "glmulti" procedure as an "exhaustive search" for this body of work.

A cross-validation procedure was implemented in this work to directly estimate the true testing error of the models being considered during the modeling process and of the final allograft TCV prediction model. The implemented cross-validation procedure allowed the testing errors to be estimated even though the healthy heart library was of moderate size (N=97). The testing error is the error of a dataset that contains no data points that were used to train the model. The importance of the testing error is it is an estimate of the model's true prediction error – as long as the testing sample is a large, realistic distribution of prediction cases. The training error is the error of a dataset that was also used to train the model but it does not necessarily represent the true prediction error. The basic issues in using the training error to assess a model's prediction capabilities are (1) it is biased to underestimate the true testing error and (2) the training error rankings do not necessarily represent the testing error rankings [118]. In particular, more flexible models, e.g., less robust models that are highly overfitted, can easily pass close to or even through many of the training data points and therefore greatly reduce, i.e., bias, the training error. This training error bias is specific to the training dataset and therefore explains the preference in using the testing error to analyze a model's performance.

Cross-validation procedures allow testing errors of relatively small datasets to be estimated by systematically separating the data into a series of training and testing datasets. The particular strengths of the cross-validation procedures when the only data available is of limited size are (1) the process reduces overestimation of the testing error and (2) the testing error variability can be averaged [118]. Overestimated testing errors can be caused by using a small training dataset, however, cross-validation allows for most of the data to be used in model training for any single fold [118]. Allowing most of the available data to train the model in a single fold helps to prevent the testing error from being overestimated. The testing error variability is a consequence of the training and testing datasets being highly variable, especially when the testing datasets is small. Cross-validation addresses the testing error variability of a small testing dataset by averaging the error of multiple folds, i.e., fits [118]. Without cross-validation techniques, appropriately large training and testing datasets are typically needed to reduce testing error overestimations and to reduce testing variability, respectively [118].

Model A was developed using a k-fold cross-validation procedure. In general, a k-fold uses $100\%\left(\frac{1}{k}\right)$ of the data for testing and $100\%\left(1-\frac{1}{k}\right)$ of the data for training at any one time in the "*k*" individual folds. Implementing a 10-fold procedure, i.e., k = 10, would result in 90% and 10% of the data being used for training and testing the model for each of the folds, respectively. Reducing the value of "*k*" results in a lower error metric variance (preferential) but at an increased testing error estimated bias (undesired) – this is known as the bias-variance trade-off [118]. A reduction in "*k*" biases the model because the training dataset becomes smaller [118]. Model B, in the next section, was developed using leave-one-out cross-validation (LOOCV). For all intents and proposes, LOOCV is a k-fold situation in which k = N, i.e., k equals the sampling size. The LOOCV process generates the lowest bias for any of the k-fold methods but at the cost of increased error variance. The reason for the higher error variance in LOOCV is a direct consequence of averaging highly correlated quantities [118]. To address the bias-variance trade-off in the initial allograft TCV prediction model, a 10-fold cross-validation process was initially chosen to focus on not biasing the modeling process while being able to estimate the testing error with a low variance.

The initial modeling process, which derived Model A, implemented k-fold cross-validation procedures in steps 4 and 7 to estimate the testing errors. The testing errors were then used to calculate the testing statistical model metric errors. As the methods, result, and discussion indicated hereto, the author took special care in selecting structural frameworks based on their testing statistical model metric errors. Testing error metrics were used herein to help ensure model selection was based on error metrics that were more reliable than the training error metrics. Furthermore, the cross-validation procedure of choice was chosen to address model error bias and variance. Step 4 presented intermediate testing statistical modeling metric error results in Tables 5.4 to 5.15 for the 8 top models to (1) help ensure error bias and variance was minimal between the top models for consideration and (2) to ensure the modeling process was appropriately on track to develop a final allograft TCV prediction model. Analysis of the cross-validation testing statistical modeling metric errors and error ranking summaries in step 7, i.e., Tables 5.28 to 5.30, were used to help select the final allograft TCV prediction model, i.e. Model A.

One important limitation of k-fold cross-validation procedures, especially for large values of "*k*", is the increased computational cost required to estimate the direct testing error. Averaging several k-fold runs helps to remove the bias of a single run further increases the computational cost, e.g., as was the case herein in which each cross-validation procedure was run 100 times. Given these facts, it was not generally reasonable to use the k-fold procedure for the exhaustive search analysis. Instead of directly estimating the testing errors from the exhaustive search procedure, there were other means available to indirectly estimate testing errors and reduced the computational cost.

The AICc metric, when calculated using the ML method, can be used as an indirect testing error metric to quickly compare the structural frameworks of interest – as was performed in step 3's exhaustive search. The various indirect testing error estimation methods, including the AICc metric, make adjustments to the training error (based on theoretical assumptions) to estimate the testing error [118]. Although it may not be readily apparent, the adjusted-R^2 is another indirect testing error measure while the R^2 does not include theoretical adjustments to the training error [118]. The key issue with the R^2 value is the value artificially improves as more terms, e.g., predictors, are added to the model – this is similar to how overfitting a model reduces the training error but does not correctly estimate the true testing error [118]. Ultimately, the AICc value and AICc difference threshold were used herein to (1) reduce computational cost and (2) quickly select a subset of models to focus on during allograft TCV prediction model development. It should be noted that because the AICc metric is an indirect estimate of the testing error the author decided this metric is not ideal for final model selection, especially because the AICc differences are less than 2 and therefore minimal. In other words, directly estimating the testing error was preferential in the modeling process when the computational cost of performing a k-fold cross-validation at a particular modeling step was reasonable, otherwise, the AICc metric was used herein.

Multicollinearity was addressed in the modeling process, via a VIF procedure, to help ensure the final allograft TCV prediction model was robust. Although mild multicollinearity will have limited effect on the model performance, high collinearity increases the estimate coefficients SEs and reduces the overall model robustness [108,109]. Relationships between the dependent and

independent variables can become unstable to the point they are no longer interpretable or even to the point that the coefficient values change signs when high multicollinearity is an issue [108,109]. Similar to overfitting a model, high multicollinearity can lower training errors without improving the testing errors [108,109]. Extrapolated predictions, i.e., predictions made outside training dataset domain, are particularly problematic for models with high multicollinearity and even those that are overfitted because of the associated decreased robustness [108]. Interpolating a prediction, i.e., making a prediction within the training dataset domain, will generally perform better – this explains in part why extrapolated prediction should be avoid.

Asking clinicians and other users of the allograft TCV prediction model, or any model, in general, to not extrapolate future predictions is advisable. Expecting model end-users to never extrapolate future predictions is unrealistic nonetheless for several reasons. First, there are perceivable clinical scenarios in which a clinical team might thing a slight extrapolation (a best-case scenario) would only increase the error of the TCV prediction only slightly and therefore be worth the clinical risk. The clinical team would likely be mistaken if multicollinearity and even overfitting issues were not addressed, i.e., if the developed model was allowed to remain unstable, because unrealistic predictions could be made. Second, even when experienced end-users avoid extrapolating predictions, there are so called “mild” extrapolations, i.e., hidden or hard to detect extrapolations, that can result in unrealistic predictions [108]. A mild extrapolation happens when a predictor is within the minimum and maximum range of two or more variables but still outside of the model’s actual training population. A generic example showing standard and mild extrapolations are illustrated in Figure 5.7. Multicollinearity was addressed herein for system redundancy, i.e., for a system safeguard, to help prevent unrealistic extrapolated allograft TCV predictions being made when an end-user performs an extrapolation intentionally or unintentionally. Future versions of the virtual HTx fit assessment tool could go a step further and include an algorithm that flags unintentional allograft TCV prediction extrapolations.

Examples of Interpolation and Extrapolation Predictions

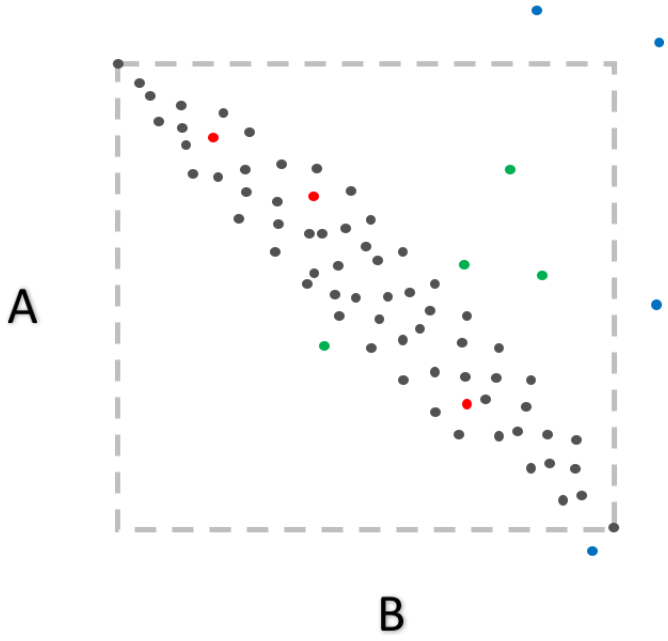

Figure 5.7: An example training data set (dark gray) was plotted against its “A” and “B” independent variables. The box (light gray) illustrated the minimum (e.g., lower left corner) and maximum (e.g., upper right corner) range of the training data. The red, green, and blue data points illustrated testing data points that would have corresponded to interpolated, mildly extrapolated, and extrapolated predictions. Notice mild extrapolations are still within the minimum and maximum ranges for the independent variables “A” and “B” but are not within the example training dataset.

Heteroscedasticity was illustrated previously in Figures 4.2 and 4.3 and confirmed with the Breusch-Pagan test in Table 5.16. The issue with heteroscedasticity in the modeling data is it increases the estimated coefficient SEs and therefore can negatively impact the model’s robustness [108]. Herein, two methods to correct for heteroscedasticity were considered: (1) weighting the estimated coefficients and (2) removing influential outliers. First, coefficient weighting used “varClass” functions with little effect on improving the models’ prediction performances or Breusch-Pagan test p-values for heteroscedasticity, as was presented in Tables 5.16 and 5.17. Second, using Cook’s distance to remove influential outliers resulted in increased model prediction performance and the Breusch-Pagan test indicated no sign of heteroscedasticity, as was presented in Tables 5.16 and 5.17. The need to address heteroscedasticity in the data was a to ensure the model was robust and help reduce unrealistic predictions, especially with mild or standard

prediction extrapolations – this is a similar reasoning to building in model redundancies/safeguards by addressing multicollinearity during model development.

During the modeling process, there were 6 statistical modeling metrics, i.e., MAE, MAPE, ME, MPE, MSE, and RMSE, and two additional metrics, i.e., sMAPE and sMPE, mainly used to assess model development and performance. These 8 metrics quantitatively highlighted unique attributes of the prediction errors, i.e., residuals, of the models being developed, validated, and considered for the final allograft TCV prediction model. In reviewing these metrics' formulas, presented in Table 5.2, it becomes apparent the large variety of metrics highlight a wide range of error information on the models being considered. Metrics that take the residual absolutes, i.e., MAE and MAPE, provide information on the overall error without considering if the error was an over- or under-prediction and without penalizing larger errors. The ME and MPE metrics function similarly to their corresponding residual absolute metrics but with over- and under-predictions considered in the metric, e.g., provide the direction of the bias, without penalization of larger errors. The MAPE and MPE metrics report relative errors by decreasing the penalty of errors for larger predictions; e.g., a 50mL error could be considered reasonable for an 800mL TCV in a teenager but not for a 75mL TCV in an infant. The MSE and RMSE squared the residuals and therefore do not provide directional over- or under-prediction information but they did penalize larger errors, regardless of the relative error. Furthermore, the RMSE takes the square root of the MSE such that the metrics unit is no longer squared, i.e., the RMSE value magnitudes are easy to interpret.

The MPE and MAPE are important metrics, in general, because they provide information on relative error, however, there is an important limitation that could even be considered a flaw with these 2 metrics [135–137]. Table 5.2 shows these two metrics' formulas are divided by the predicted value and this results in these metrics applying larger penalties on over-predictions. For example, if the MAPE metric had an under-prediction of 50% (i.e., Actual = 2 and Predicted = 1) then switching the actual and predicted values would result in an over-prediction of 100% (i.e., Actual = 1 and Predicted = 2). This higher penalty for over-predictions could bias the developed model to under-predict. This asymmetric bias of the MAPE and MPE metrics can be corrected, i.e., made symmetrical, by replacing the denominator with an average of both the actual and predicted

values – these are known as the sMPE and sMAPE metrics in Table 5.2 [136]. Although the MPE and MAPE metrics were used throughout the modeling process the final model selection included the sMPE and sMAPE metrics, i.e., Table 5.30, to help identify the “best” model.

The importance of reporting multiple performance metrics, in summary, is it helps to better illustrate how a prediction model performs overall. Not every metric shows directional bias of a model, e.g., MAE, MAPE, MSE, or RMSE. Other metrics can show directional bias, e.g., ME and MPE, but cannot differentiate a net-zero directional bias vs. a model that has perfect prediction. Some metrics penalize the relative error size, e.g., MPE, MAPE, sMPE, and sMAPE, while other metrics penalize larger errors, regardless of relative size, e.g., MSE and RMSE. Furthermore, there are metrics that provide insight into a model while sacrificing or biasing another attribute that may not be ideal or even problematic in interpreting another attribute of a model’s performance, e.g., MAPE and MPE. The modeling process herein considered a wide range of statistical modeling metrics to understand the top model performances so a “best” model could be identified in the final model selection process.

The testing statistical model metric rankings demonstrated Model 2 outperformed with the MAE, ME, MSE and RMSE metrics in Table 5.29. The MAE and ME showed Model 2 had the lowest absolute testing error and the lowest over- or under-prediction bias, respectively. The MSE and RMSE showed Model 2 was less likely to produce larger errors relative to other models. The MAPE and MPE showed a slight and large decrease in the performance of Model 2, however, after addressing the asymmetric of these metrics in Table 5.30, the sMAPE and sMPE showed large improvements in Model 2. The final results in Tables 5.29 and 5.30, after the symmetry correction, demonstrated Model 2 was ranked 1st for MAE, sMAPE, ME, MSE, and RMSE and was ranked 2nd for sMPE. The MAPE and MPE metrics were disregarded in the final assessment due to the well-established symmetry issues and given the sMAPE and sMPE helped to distinguish Model 3 for 3rd place in model performance when it was previously unclear if Model 3 outperformed Model 8. The testing statistical model metric results demonstrated Model 2 was the “best” model and therefore became Model A.

Final Model A TCV prediction results that were used to calculate the final statistical modeling metric results that were presented in Figures 5.3-6. These residual results generally fluctuate about the line of unity in Figures 5.3-4 with a slight bias of under-prediction for the largest of heart predictions. This suggest Model A biases under-predictions for the largest of hearts but the relative errors are still within the more generally visible ±25% bandwidth in Figure 5.4. The Shapiro-Wilk test results (p-value ≤ 0.0001) and Figure 5.5 indicate the residuals were not normally distributed. However, visual inspection of Figure 5.6 appears to be suggesting the residuals are somewhat near normality, centered nearly at 0, and generally suggested Model A performed well. Figure 5.6 does suggest errors between 0 and 50mL was the most frequent error size.

The development of Model A was driven by the AICc difference profiles in the initial exhaustive search, the final estimated coefficient p-values, and the comparison of the statistical modeling metrics throughout the modeling process. Age was also considered initially, however, from a practical point-of-view Age was excluded early in the modeling process. The practical issue was getting donor Age in months was difficult for older donors while Age in years was too large for infants. Fortunately, the inclusion or exclusion of Age had no effect on the final predictors, i.e., Sex, Ht, and BMI, used in the *log-log* transform modeling process which was the datatype used to develop Model A. It is worth noting that non-parametric techniques were not considered in the modeling process because (1) they do not generate an equation but, generally speaking, are a set of smoothing criteria that require access to the training data and (2) they would have required a much larger dataset to train the model [118,138]. Although the author developed Model A as presented herein, it was suggested the initial modeling process could be improved upon and therefore Model B was developed. The development and results of Model B were presented and compared to Model A in the remaining sections of chapter 5.

5.4 Methods and Materials of the Improved Allograft TCV prediction Models

The “improved” allograft TCV prediction model, i.e., Model B, was developed after reviewing the initial modeling process methods and preliminary results. Seven key modifications to the modeling

process were made. First, the VIF procedure was performed in an iterative process but the highest VIF predictor was not automatically excluded. Instead, predictors with large VIF values were kept in the modeling process if they were shown to have strong correlation with TCV, however, predictors with lesser VIF values were removed. The VIF procedure removed predictors until all remaining predictors had VIF values that dropped enough to meet the VIF threshold. Second, the VIF cutoff threshold was increased (i.e., < 20). Third, the AICc difference cutoff threshold was increased (i.e., ≤ 3) – models that meet this criterion lie somewhere between having some and substantial chance of having equivalent performance [124]. Fourth, the k-fold cross-validation was changed to a leave-one-out cross-validation (LOOCV) procedure. Although Model A originally used 10-fold cross-validation to estimate the testing errors during model development and validation, Model A's testing errors were recalculated using LOOCV in the next section to compare the initial model results to Model B. Fifth, the predictors were standardized for coefficient estimation. Sixth, the coefficients were estimated with an iteratively reweighted least squares method based on the Huber weight function to ensure the developed model was robust to heteroscedasticity and outliers [139]. Seventh, estimated coefficients and their corresponding bootstrap-derived confidence intervals were investigated to ensure statistical significance.

Prediction errors are a reality for any statistical model even though care is taken to minimize these errors. A key intent of the virtual fit assessment tool was to help clinicians safely expand a patient's donor pool by maximizing allograft TCVs that could be accepted, however, under-predictions were clinically concerning when the oversized allograft limits are being pushed. This opinion that over-predictions were clinically preferable to under-predictions led to the development of a secondary 0.75 quantile regression model, i.e., Model B*. Model B* was developed from Model B's structural framework with modifications that penalized under-predictions 3 times more.

5.5 Results of the Improved Allograft TCV prediction Models

Models B and B* were developed with Age automatically excluded in the modeling process with the logic used during the development of Model A. The final, improved allograft TCV models were

presented in Table 5.31 in which they predicted allograft TCVs in mL. All continuous healthy heart library variables, including Age, were standardized and included in Table 5.31. The inputs for the standardized TCV, Age, Ht, Wt, BSA, and BMI variables were in mL, months, cm, kg, m^2, and kg/m^2, respectively. The standardized TCV was not used in the modeling process but was included to provide the variable's mean and standard deviation. The standardized Age was included because it was originally considered in the modeling process until it was realized that getting older donors' ages in months was impractical for the virtual fit assessment tool. Sex was a dummy variable in which male was numerically code as "1" and female was coded as "0". Sex was not standardized because it was a nominal, categorical variable.

Models A and B were estimated using least squares regression techniques in which the model estimates the conditional mean [140]. Model B* was not a least squares regression because it aimed at predicting the conditional 0.75-quantile of the response variable [140]. The 0.75-quantile regression model was fitted such that 75% of the training set was below the expected fit. If the training data represented the real-world than this fit would represent 75% of predictions would be over-estimated.

Final Results for Improved Models' B and B*

Model B and B*							
$pTCV_{\text{Model-B}} = exp\left(\boldsymbol{\beta_1} + \boldsymbol{\beta_2} * \left[\frac{\ln(Ht) - 4.963}{0.260}\right] + \left(\boldsymbol{\beta_3} + \boldsymbol{\beta_4} * \left[\frac{\ln(Ht) - 4.963}{0.260}\right]\right) * \left[\frac{\ln(Wt) - 3.795}{0.714}\right] + \boldsymbol{\beta_5} * Gender + \frac{\boldsymbol{\beta_6}^2}{2}\right)$							
Parameter	**1**	**2**	**3**	**4**	**5**	**6**
β	**6.053** (0.027) [-224.2]	**0.340** (0.069) [-4.9]	**0.302** (0.057) [-5.3]	**0.016** (0.020) [-0.8]	**0.099** (0.032) [-3.1]	**0.132** (0.012) [-11.0]
β*	**6.137** (0.041) [-149.7]	**0.282** (0.107) [-2.6]	**0.338** (0.084) [-4.0]	**0.015** (0.028) [-0.5]	**0.125** (0.028) [-4.5]	**0.000**
Developed Standardization Equations for mTCV and Parameters						
mTCV	**Gender**	**Age**	**Height**	**Weight**	**BMI**	**BSA**
$\frac{\ln(mTCV) - 6.132}{0.621}$	$Gender$	$\frac{\ln(Age) - 4.840}{0.849}$	$\frac{\ln(Ht) - 4.963}{0.260}$	$\frac{\ln(Wt) - 3.795}{0.714}$	$\frac{\ln(BMI) - 3.079}{0.295}$	$\frac{\ln(BSA) - 0.284}{0.481}$

Table 5.31: Model B and B* estimated coefficients (bolded), SEs (in parenthesis), and t-values (in brackets) were presented in the **β** and **β*** rows; i.e., placing the **β*** rows estimated coefficients into Model B would be Model B*. Both models used the exact same structural framework but their estimated coefficients and SEs varied. The model suggested an individual's Ht and Wt increases with TCV. The model also suggested TCVs were larger in males. The interaction term of Ht with Wt suggested Wt had a larger, positive effect on the TCVs of taller individuals.

The LOOCV modeling statistical test results for Models A, B, and B* were presented in Table 5.32 using all the health heart library data points, i.e., N = 97. The testing ME was the only metric Model B outperformed Model A, otherwise, Model A outperformed Model B. Model B*'s

testing statistical modeling metrics indicated it had the worst performance out of the 3 reported models; however, this was not surprising due to the over-prediction bias that was forced in the 0.75-quantile regression process. The LOOCV pTCV results for Models A, B, and B* were presented in Figures 5.8 and 5.9 as actual vs. observed and relative error plots. The plots in Figures 5.8 and 5.9 further supported the testing statistical model metric results in Table 5.32 that Model B did not outperform Model A, in general. Figure 5.9 might have further suggested that Model A outperformed Model B by having a lower absolute relative error for the smallest of TCVs.

Testing Statistical Metric Results for Initial and Improved Models

Model	*MAE*	*MAPE*	*ME*	*MPE*	*MSE*	*RMSE*	*sMAPE*	*sMPE*
A	63.5mL	12.1%	5.5mL	-1.0%	8408mL^2	91.7mL	12.0%	0.2%
B	64.0mL	12.7%	1.2mL	-2.3%	8707mL^2	93.3mL	12.3%	-1.0%
B*	74.2mL	16.1%	-46.4mL	-11.8%	10606mL^2	103.0mL	14.4%	-9.8%
Standard Deviation for Metrics' Means								
A	66.5mL	10.0%	92.0mL	15.7%	18586mL^2	136.3mL	2.4%	3.8%
B	68.2mL	10.8%	93.8mL	16.6%	20377mL^2	142.7mL	2.5%	4.0%

Table 5.32: The LOOCV modeling statistical metric results between the initial and "improved" processes. Model A outperformed Model B in 7 of the 8 testing statistical model metrics – the ME result was the exception. Although the testing statistical model metrics indicated Model B* was the worst model, this decrease in metric performance was expected because of the model was made to bias over-predictions, i.e., set to offset 75% of the training data. Model A's standard deviations were smaller for all 8 testing metrics.

Model A, B, and B*'s pTCV vs. the mTCV

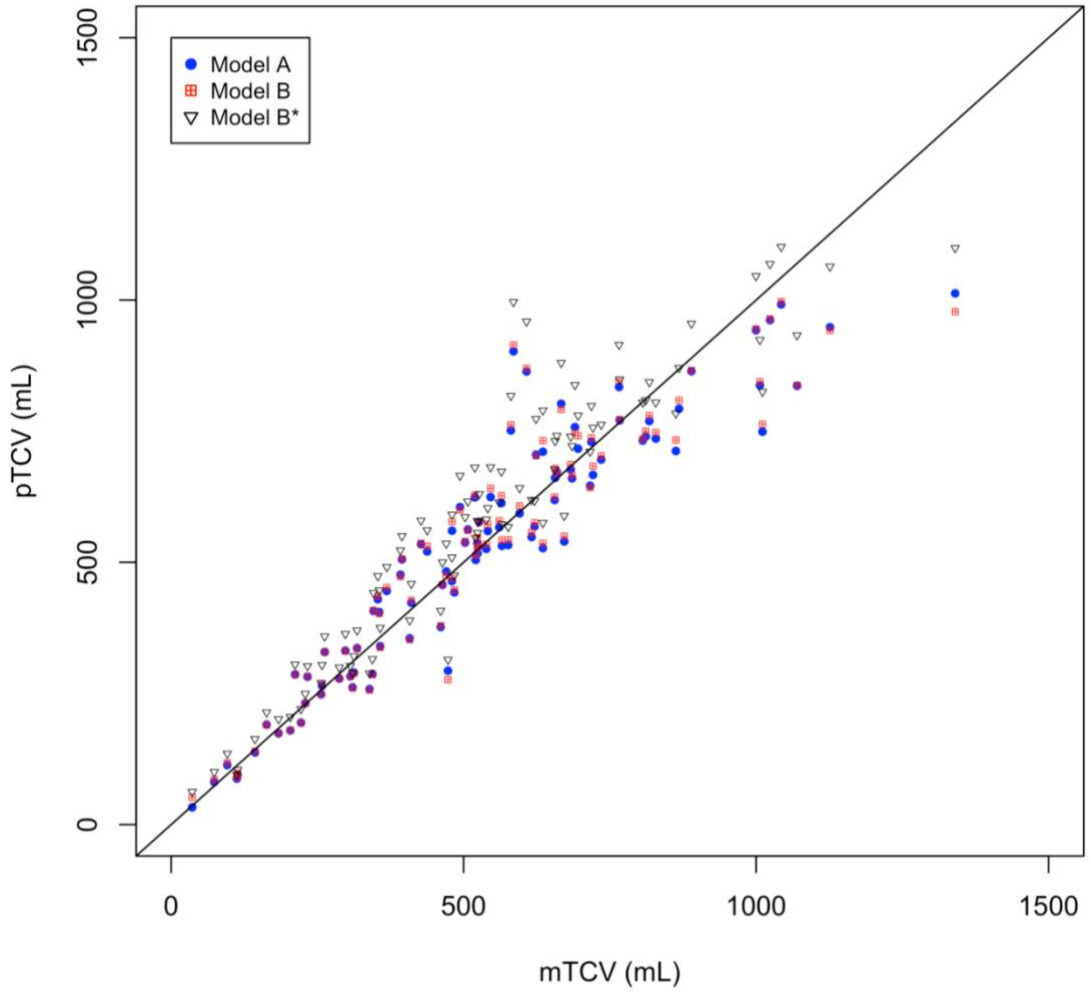

Figure 5.8: The pTCV vs. mTCV plot demonstrated Models' A, B, and B* LOOCV pTCV fits about the slope of unity (black line). Data points above the slope of unity were over-predictions. Notice, B* has an offset that pushed every B result upwards; this was a result of the model's over-prediction bias. Visual inspection of the current plot illustrated no obvious deviations between Models A and B.

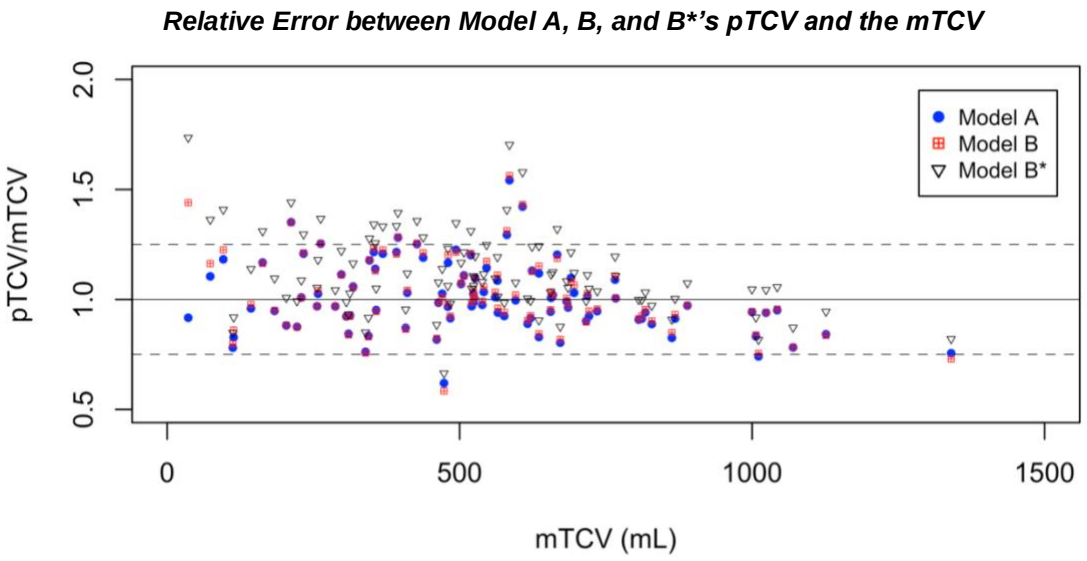

Figure 5.9: The pTCV vs. mTCV plot demonstrated Models' A, B, and B* relative LOOCV pTCV fits to the horizontal fit of unity (black line). The dashed lines illustrate the 0.75 and 1.25 relative error levels. Notice, B* had an offset that pushed every B result upward due to the over-prediction bias. Visual inspection of the current plot illustrated no obvious deviations between Models A and B, in general, except for the smallest TCVs.

Figures 5.10 and 5.11 presented plots to compare the LOOCV pTCVs and the relative LOOCV errors between Models A and B. A tight fit about unity was illustrated in Figure 5.10's Model A vs. Model B plot. The tight fit suggested, in general, Models A and B were nearly consistent in prediction performance even though Model A outperformed Model B based on the LOOCV modeling statistical metric results. Furthermore, the relative error results illustrated in Figure 5.11 suggested that for all but the smallest of TCVs the models were nearly consistent in prediction performance. An ANOVA test failed to reject the null hypothesis that the Models A and B's relative error predictions were statistically the same (p-value = 0.5588, Cohen's distance effect size = 0.08412). The small effect size confirmed the p-value finding, i.e., failure to reject the null hypothesis, was also supported in a practical sense. In other words, the p-value and effect size findings indicate the models' relative prediction performances were equivalent in the practical sense. Zoomed in views of Figures 5.8 and 5.9, focusing on the lower TCV range for Models A and B, were presented in Figures 5.12 and 5.13. Figures 5.11, 5.12, and 5.13 illustrated although

Models A and B fluctuated about the slope of unity, Model B consistently made larger predictions than Model A for mTCVs up to 150mL.

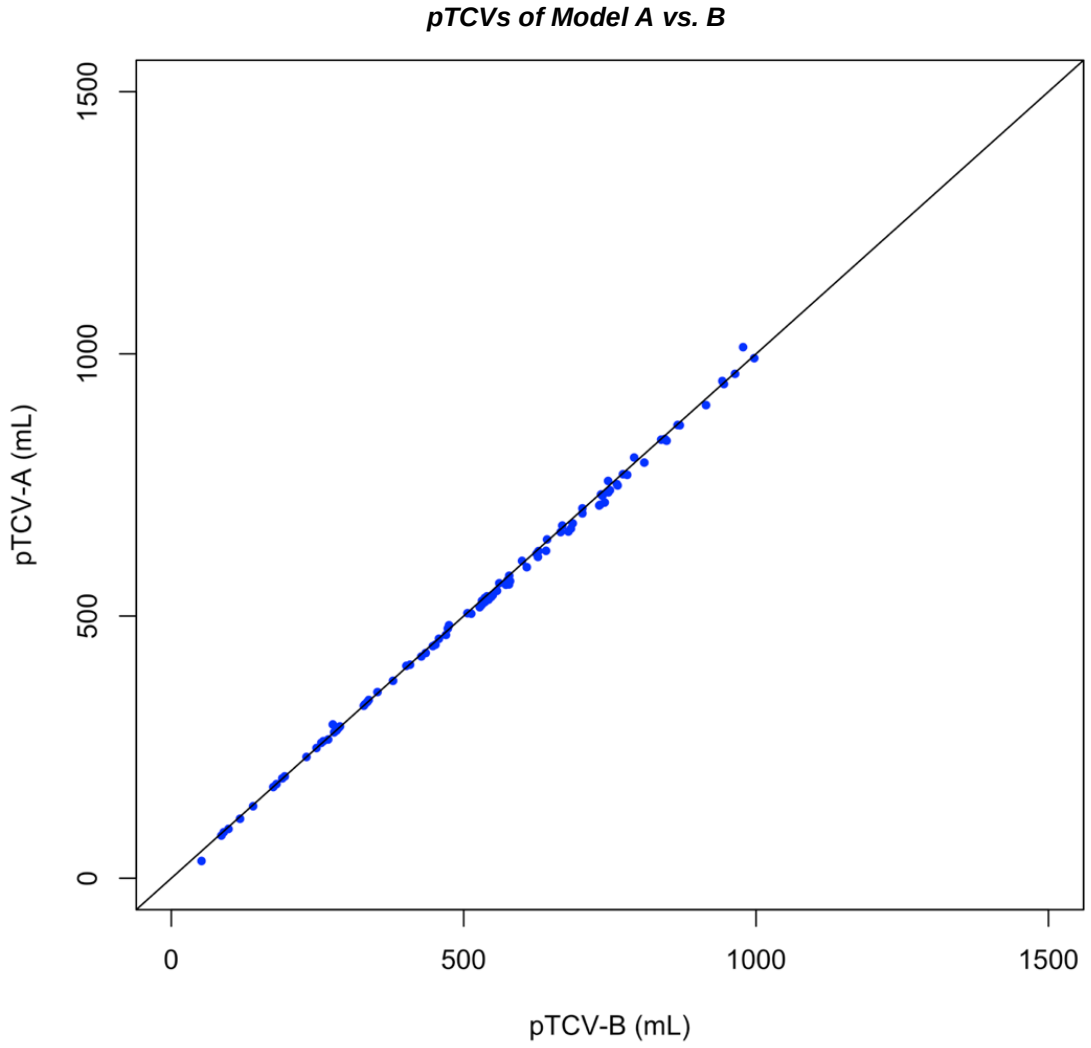

Figure 5.10: Plot illustratively comparing the pTCV results between Models A and B. The plot's results illustrated a fit tightly around unity and therefore suggested the models' predictive performances were similar even though the testing statistical modeling metrics indicated Model A outperformed Model B. Model B pTCV results were chosen to be the reference data along the x-axis because it was the final model used in the virtual heart transplant fit assessment tool.

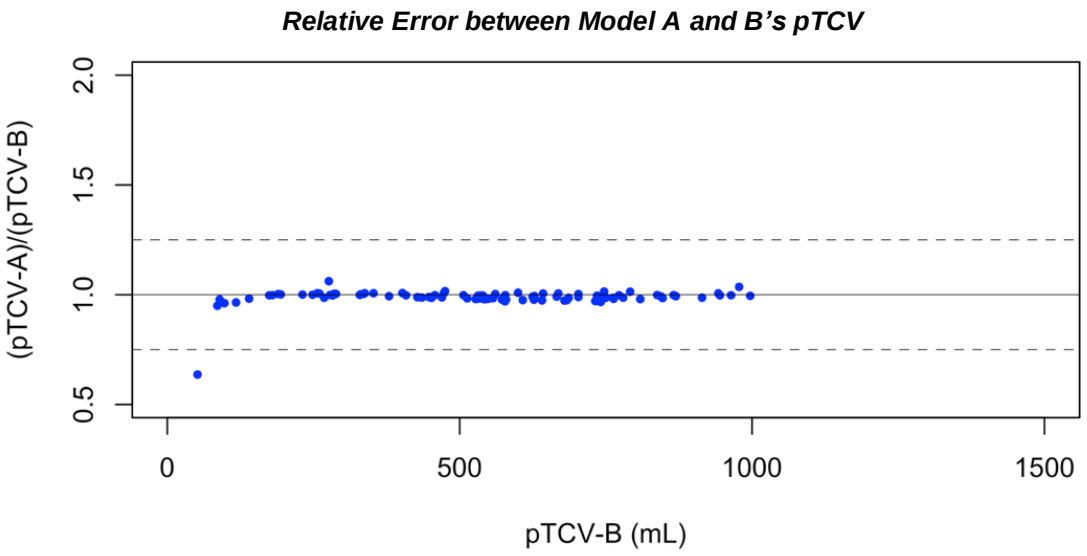

Figure 5.11: Comparative plot showing relative errors between Models A and B. The generally tight fit of relative error around unity further supported both models make similar predictions for TCV – excluding the smallest of TCV predictions. For the smallest of pTCVs, Model B made larger predictions than Model A. Model B pTCV results were chosen to be the reference data along the x-axis because it was the final model used in the virtual heart transplant fit assessment tool.

Model A and B's pTCV vs. the mTCV: Zoomed In

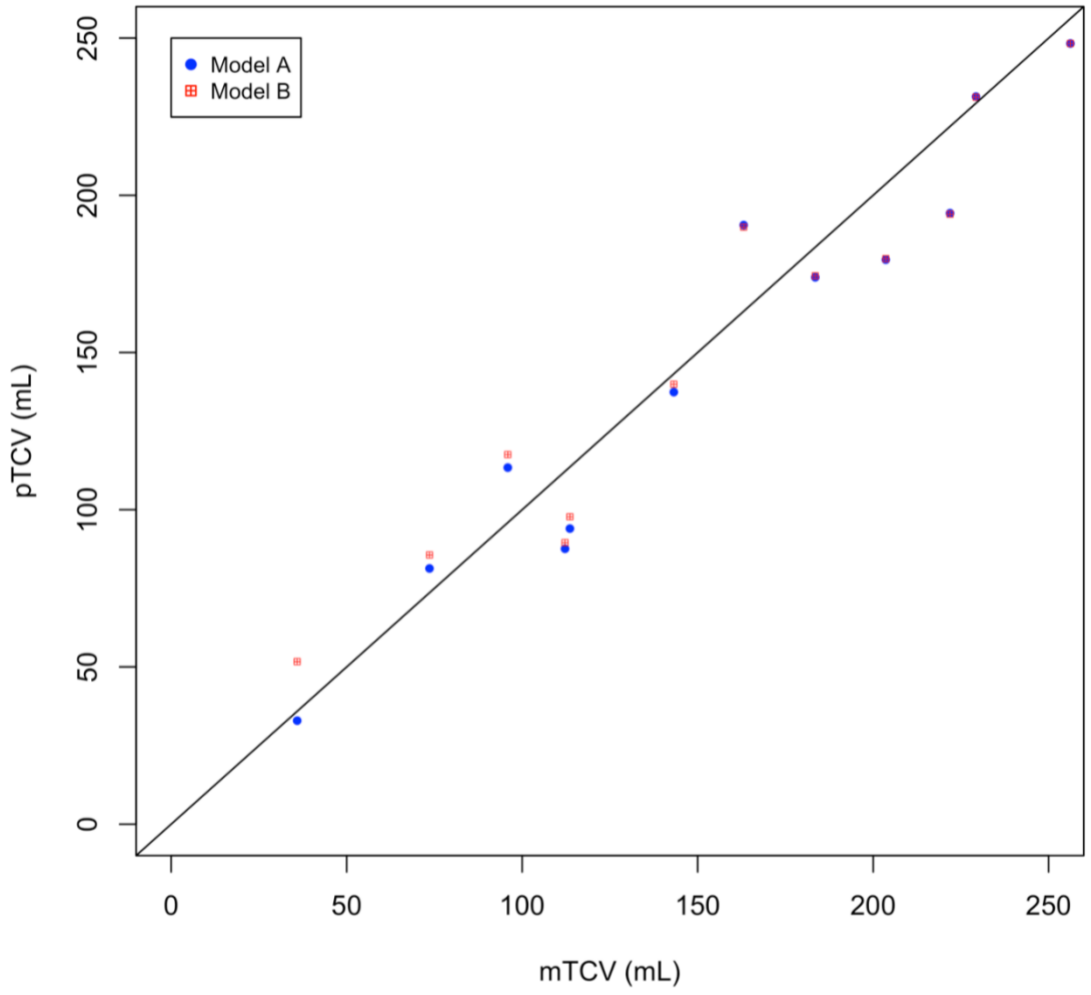

Figure 5.12: Zoom in of the pTCV vs. mTCV graft in Figure 5.4 illustrated a subtle, systematic bias between Model A and B's predictions for TCVs ≤ 150mL. However, for TCVs ≤ 150mL, both models still fitted about the slope of unity (black line). Although there was a systematic difference in the pTCVs between the models, the models did not suggest a bias between the pTCVs and the mTCVs.

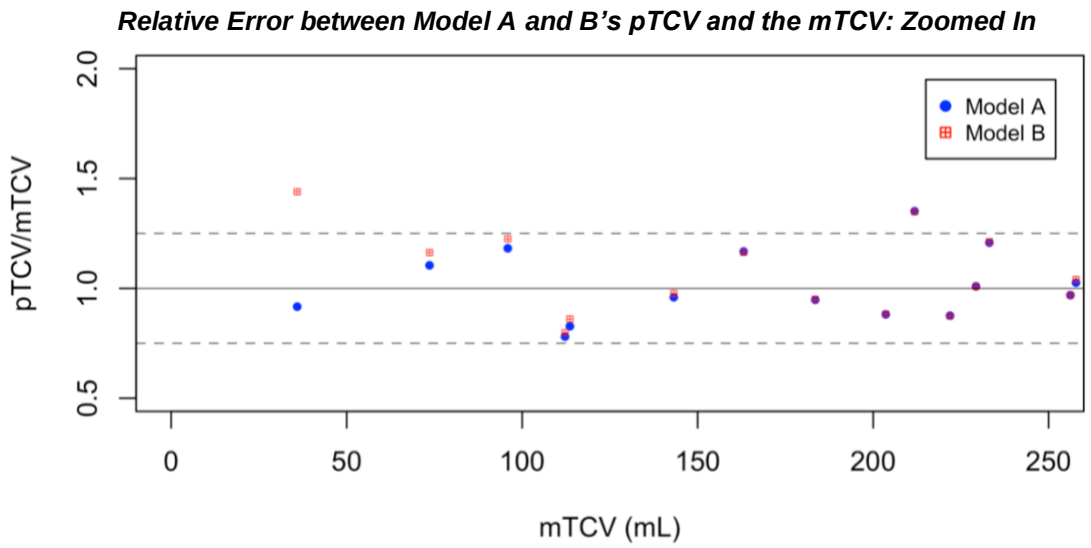

Figure 5.13: Zoom in of the relative error plot illustrated a systematic difference between Models A and B's predictive performances for TCVs ≤ 150mL.

Model B's pTCV residual distribution was presented in Figures 5.14 and 5.15 with both a quantile-quantile plot and histogram plot. Although visual inspection of the histogram might be interpreted as being a normal distribution, the normal quantile-quantile plot presented "tails" that appeared to be more than negligible. The null hypothesis of normality was rejected (p-value ≤ 0.0001) with a Shapiro-Wilk test and therefore confirms the errors were not normally distributed. Visual inspection of the quantile-quantile plot and histogram suggested the deviation from normality was mild and therefore suggested the lack of normality was unlikely to negatively affect the overall Model B performance.

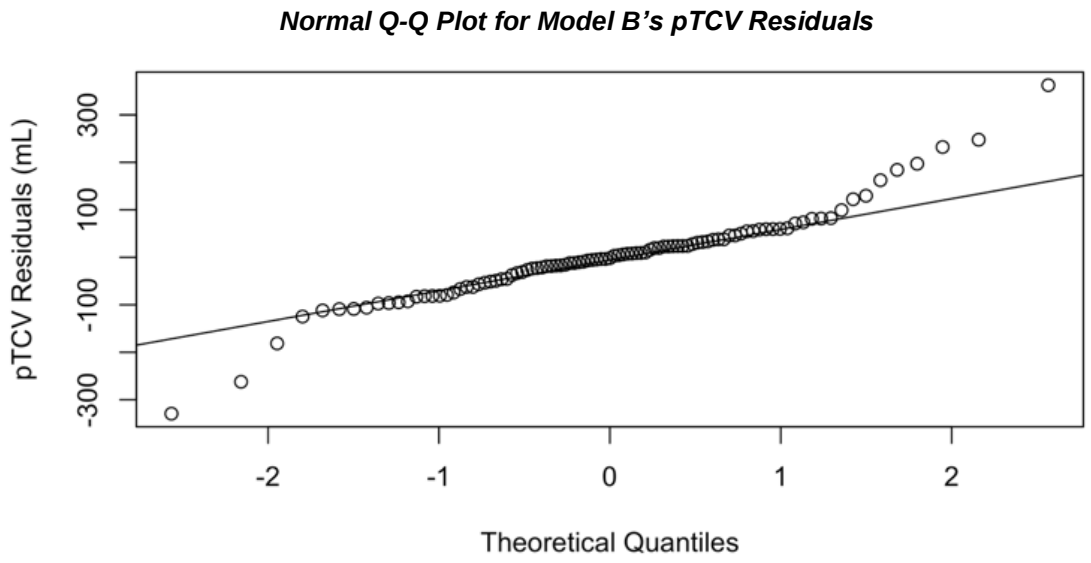

Figure 5.14: A normal Quantiles-Quantiles plot was presented to investigate the normality assumption of the Model B pTCV residuals. These plots suggest the increased likelihood of normality as the residuals better fit a single, linear line. The residuals herein generally fitted a linear line with the exception of visually appearing mild "tails" to suggested a near-normal distribution was present. A Shapiro-Wilk test rejected the null hypothesis of normality (p-value ≤ 0.0001) to further confirm the lack of normality.

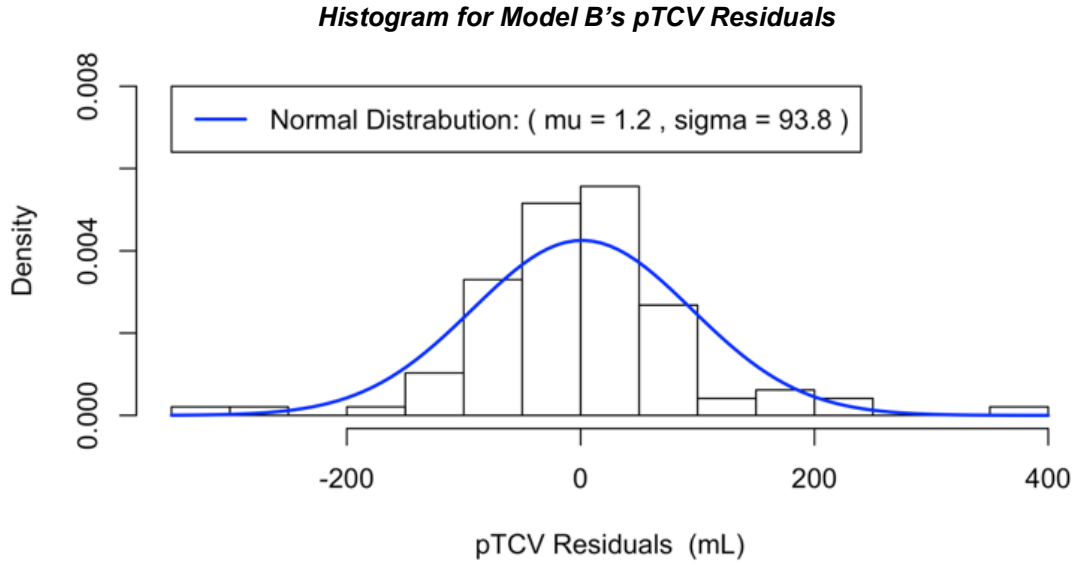

Figure 5.15: A frequency plot was created with the Model B pTCV residuals placed in one of 14 histogram containers. Visual inspection of the plot suggested either the residuals had a normal or near-normal distribution. The continuous curve (blue) was included to present an actual normal distribution with a mean (1.2mL) and standard deviation (93.8mL) matching the residual data. A Shapiro-Wilk test rejected the null hypothesis of normality (p-value ≤ 0.0001).

5.6 Discussion of the Improved Allograft TCV prediction Models

Model B was developed with the intent to build upon and improve Model A's modeling protocol. The predictors were standardized in Model B to ease coefficient comparison but did not improve the overall model's predictive capability [108]. The LOOCV testing statistical modeling metrics in Table 5.32 suggested Model A generally outperformed Model B – the ME was the only metric in which Model B outperformed Model A. Furthermore, the St. Dev. values were smaller for Model A in Table 5.32. Visual analysis of Figures 5.8 to 5.13 suggested that, although Model A outperformed Model B, their prediction performances were generally similar in the practical sense – the only obvious but minor exception was for the smallest of infant TCVs. A closer visual inspection of Figure 5.13 further indicated Model A was particularly successful at infant TCV predictions ≤ 100mL – relative to Model B. An ANOVA test comparing the models' relative errors further supported the visual graph inspections that Models A and B's prediction performances were similar for all practical purposes (p-value = 0.5588, Cohen's distance effect size = 0.08412).

The residual distributions for Models and A and B were not found to be normally distributed; i.e., the null hypothesizes of normality were rejected using the Shapiro-Wilk test (p-value = 0.0002 and p-value ≤ 0.0001, respectively). The slightly larger p-value for Model A and the visual comparisons of Figures 5.5 and 5.6 to Figures 5.14 and 5.15, respectively, suggested Model A was relatively (but minimally) closer to being normally distributed. It is worth noting that although the p-value difference for the Shapiro-Wilk tests was measureable, it was also practically negligible. Nevertheless, visual inspection of the histograms in Figures 5.6 and 5.15 suggested the models were near-normally distributed for all practical purposes and therefore were unlikely to negatively affect the models' practical performances [108]. The previous observation that Model B had an ME metric result that outperformed Model A, i.e., that had an error mean that was closer to zero, was further illustrated when comparing the histograms. The residual histogram plots suggested the improved ME metric performance of Model B came at the cost of a slightly larger variance.

The comparison of Models A and B indicated Model A slightly - but constantly for the majority of testing statistical modeling metrics and secondary analysis methods - outperformed

Model B. However, visual inspection of Figures 5.8 to 5.13 and an ANOVA test of the relative errors suggested there was no obvious practical predictive performance differences between the models. Furthermore, there was no obvious practical predictive capability differences between the models for the smallest of TCVs, in general. The only concerning large relative prediction error in Figure 5.13 (for Model B) was for the smallest heart in the healthy heart library, however, this prediction was an extrapolation due to how the LOOCV procedure is implemented. Large extrapolation errors, especially if the model is not perfectly robust, can produce wild predictions and therefore this large prediction discrepancy should not be overly concerning.

The testing statistical modeling metrics indicated Model A was generally better at prediction performance. This better prediction performance of Model A might suggest the model should be used in the virtual heart transplant fit assessment tool – not Model B. However, the prediction performance of Model B was practically the same to Model A, i.e., the difference in performances were found to be statistically-insignificant. Ultimately, Model B was the final model used in this body of work for the virtual HTx fit assessment work. The reasons Model B was the final model used in this work was due to (1) a limitation in the preliminary Model A development and (2) the author originally thought the SEs between the models were comparable. The SE values were not comparable but the t-values were. The coefficient t-value magnitude sizes were inconsistent between the models, however, the St. Dev. metric values in Table 5.32 were consistently smaller for Model A to suggest it was a more robust model.

Before, during, and after the development of Model B, the Model B results were compared to the preliminary Model A results. The completion of Model A had been put on hold while Model B was being developed – it was only during the finalization of Model A for being included in this thesis that a few issues were highlighted. The author was either not aware of these issues or not aware of the consequences of these issues would have on data interpretation at the time Model A was put on hold during the development of Model B. First, while the initial step 4 modeling process analyzed testing error data results (from a k-fold cross-validation process), the step 7 modeling process analyzed training error data results. At the time of Model B develop, the author though the testing error data from the initial k-fold procedure at step 4 was enough. Second, the initial model

process used k-fold cross-validation and then the “improved” method, i.e., Model B, used LOOCV – there is a well-known bias-variance trade-off between these validation methods that can have impact on the final results [118]. Third, the author did not originally implement a true k-fold cross-validation in step 4. The issue with the original k-fold procedure not being a true k-fold was the author’s computer script did not consistently separate the healthy heart library, for each of the folds, into sets of 10% and 90% for the testing and training datasets. Failing to appropriately divide the dataset into training and testing subsets was a consequence of how the *R* “set.seed” function was implemented – examples using the seed function were presented in the textbook by *James et al.* [118]. In fact, there were scenarios in which the testing and training datasets contained 0% and 100% of the testing dataset.

The preliminary initial modeling process (i.e., the process that developed Model A) was corrected by (1) correctly using k-fold testing error results in step 7 (during model selection of the initial modeling process) and (2) fixing the *R* “set.seed” function implementation so a 10% and 90% data split was consistent. For the final analysis comparison of Models A and B the author used final LOOCV testing errors from step 7; again, Model A used k-fold during the final model selection before the models were compared. To get the LOOCV testing errors for Model A, the author had to re-calculated step 7’s testing errors – this explains the testing error differences between Tables 5.28 and 5.32. This thesis contains the finalized, corrected Model A results herein and the flawed, preliminary Model A results were removed.

The limitations in the preliminary initial modeling process helped to derive Model B. Once the issues in the preliminary modeling process were observe they were corrected for and demonstrated Model A generally outperformed Model B based on the testing statistical model metrics. However, the better performance of Model A was not statistically-significant (i.e., the ANOVA test comparing the models’ relative errors). Given:

1. the performance between Model’s A and B were statistically similar,

2. a more conservative Model B* was already developed, and

3. the virtual HTx fit assessments were already performed with surgeons by the time the issue with preliminary Model A process was discovered,

the author continued to use Model B.

CHAPTER 6

ANALYSIS OF THE VIRTUAL HEART TRANSPLANT FIT ASSESSMENT TOOL'S CLINICAL UTILITY

The novel virtual HTx fit assessment tool was developed with the intent to help clinicians safely expand their patients' donor pools. The clinical utility of the tool was investigated under 2 main retrospective simulated virtual HTx fit assessment studies. The studies were designed based on the historical data that was available to start answering the following 2 important questions:

1. Can the tool help clinicians expand their patients' donor pools?

2. Can the tool help clinicians perceive fit-related complications?

In the 1st simulation study, surgeons identified the maximum allograft they were willing to take using the virtual tool. In the 2nd simulation study, surgeons virtually HTxed analogue allografts that were identified with the actual donors' parameters and assessed for perceive fit-related complications. Additionally, a unique case study in which actual donor images were available was also discussed.

Chapter 6 covers the simulated virtual HTx studies and their corresponding findings. First, the chapter covers the generalized virtual HTx fit assessment procedure and the studies' underlining methods. Second, the results were analyzed and conclusions on the tool's clinical utility were presented. Third, a case study in which a virtual HTx was performed using donor images was discussed and related to the simulation studies' results. For this body of work, the analysis of the novel tool's clinical utility constituted as Aim 3.

6.1 Methods and Materials for Virtual Heart Transplant Fit Assessments

Virtual HTx fit assessments were performed in Mimics software. The virtual assessments consisted of a surgeon strategically fusing an allograft geometry into a patient's CT or MR image. Allograft placement was performed in real-time by translating and rotating the geometry into place while viewing the 3 orthogonal planes, i.e., axial, coronal, and sagittal. The surgeon assessed for perceived oversized allograft fit-related complications during and after the virtual fit procedure. The surgeon needed to consider the allograft's potential effect on the local anatomy and physiology after final placement. Considerable allograft overlay onto critical structures (e.g., aorta, airway, and diaphragm) and/or rigid structures (e.g., ribs, sternum, and vertebral column) were perceived to be particularly indicative for a potential poor outcome. In general, healthy heart library reconstructions served as allograft analogues for the virtual assessments. Donor images were required for a virtual assessment to be performed using an actual donor's reconstruction and not rely on a healthy allograft analogue – a unique case study using donor images was available and was included at the end of chapter 6. Surgeons were specifically blinded to patient and actual donor body weights and to clinical outcomes. Studies 1 and 2's virtual fit assessments were retrospective, simulated procedures.

6.2 Methods and Materials for Study 1: Expanding Upper Donor Pool Ranges

Simulated maximum allograft virtual HTx fit assessments were performed on listed PCH HTx patients with pre-transplant CT or MR images. The study's objective was to see if the tool resulted in clinicians accepting traditionally perceived oversized allografts, i.e., accepting donors with traditionally larger than normal body weights, during the virtual assessments. There were 45 PCH patients that were pulled, de-identified, and analyzed for research in a retrospective chart review that spanned 2010 to 2015. The patients in the current study were carefully selected so they could be reused in study 2. Additionally, all available national level patients spanning 2010 to 2015 were

pulled from the Pediatric Heart Transplant Society (PHTS) and the United Network of Organ Sharing (UNOS) datasets for comparison.

Listed patient TCVs were reconstructed from pre-operative CT or MR images, using Mimics, and then measured. Surgeons used the patient's native mTCV to select an initial allograft reconstruction from the healthy heart library for simulated HTx consideration in a virtual heart transplant fit assessment. Surgeons assessed the fits and then, in an iterative fashion, selected another healthy heart reconstruction for review by intentionally increasing or decreasing the next allograft's mTCV. The iterative process was repeated until the surgeon identified the maximum allograft TCV they were willing to accept based on perceived fit-related complications. Once the surgeon identified the maximum allograft TCV they were willing to accept, the iterative fit assessment was ended and the maximum mTCV value was recorded. In summary, the virtual transplant and assessment followed the general procedure (previously discussed) but with the objective to identify the maximum allograft size the surgeon would accept. In identifying the maximum allograft, surgeons were asked to choose the largest allograft within reason, e.g., no "last-ditch effort" scenario for fear this would be the last allograft the patient would likely be offered before succumbing to heart failure. The surgeon was blinded to the actual transplant cases.

The maximum DRBW ratio was calculated using the patient and maximum donor's, i.e., using the selected healthy heart patient's, measured Wt. Simulated and national database DRBW ratios were visually and statistically analyzed. The DRBW ratio analysis included comparisons of the virtual results with actual and upper listed DRBW ratios. Analysis included a series of ANOVA tests (Null hypotheses = 0, Alpha = 0.05) and Student T-test to compare the DRBW ratio results. The Student T-test were performed in JMP as a series of two mean ANOVA comparisons.

6.3 Methods and Materials for Study 2: Perceiving Fit-Related Complications

Simulated allograft virtual HTx fit assessments were performed on pre-transplant CT or MR images of PCH patients that did have HTxs with actual clinical outcomes. The study's objective was to see if the tool resulted in clinicians perceiving both appropriate fits that were void of over-sized fit-related

complications and oversized fits that resulted in over-sized fit-related complications. The PCH patients in the current study were the same patients in study 1, i.e., the expanding donor pools study, because to undergo a HTx required these individuals to be listed such that they could acquire a donor offer.

Actual donor parameters and Model B were used to determine donor pTCVs. The donor's pTCV was used to select an allograft reconstruction analogue from the healthy heart library for the virtual fit assessment. A virtual assessment was performed and the surgeon's fit-related observations were recorded, e.g., concerns for or lack of concerns for fit-related complication. The surgeon was blinded to the actual transplant cases.

The surgeon's perceived findings were than compared to actual patient fit-related outcomes. Outcomes known to or clinically perceived to be associated with allograft overfitting were analyzed – these outcomes included pathological symptoms, e.g., pulmonary vein stenosis, and post-operative complications, e.g., delayed sternal closure. Perceived and actual outcome frequencies were analyzed and presented. Analysis included a series of chi square test to determine if the frequencies were statistically-significant.

6.4 Results for Study 1: Expanding Upper Donor Pool Ranges

The virtual maximum DRBW ratios (from the maximum TCVs surgeons were willing to take using the virtual fit assessment tool) were compared to the accepted and upper listed DRBW ratios at both the local, i.e., PCH, and national levels. A preliminary series of ANOVA tests were performed and demonstrated there were statistically-significant differences between (1) the PCH and the tool's DRBW ratio means and (2) the national level and the tool's DRBW ratio means. The statistically-significant ANOVA results justified the need to perform a series of 5 Student T-test that compared specific DRBW ratio population pairing means. The 5 Student T-test DRBW ratio mean comparisons were presented in Table 6.1 and were referred herein as *C1*, *C2*, …, and *C5* in which "*C*" stands for "comparison". The PCH and virtual maximum quantile distributions were presented pictorially and quantitatively in Figure 6.1 and in Table 6.2, respectively. The national level and

virtual maximum quantile distributions were presented pictorially and quantitatively in Figure 6.2 and in Table 6.3, respectively.

The PCH results found the hospital's accepted allograft DRBW ratio mean was smaller than what surgeons took using the tool but it was not statistically-significant (*C1*; ratio mean of 1.68 vs. 1.94; p-value = 0.0725; effect size = 0.39). The 0.39 effect size suggested there was between a 78.7% and 72.6% overlap between PCH's accepted and the tool's DRBW ratio sample population distributions [113,141]. The PCH results found the hospital's upper DRBW ratio listing mean was statistically larger than what surgeons took using the tool (*C2*; ratio mean of 2.90 vs. 1.94; p-value < 0.0001; effect size = 1.05). The 1.05 effect size suggested there was between a 44.6% and 41.1% overlap between PCH's upper listed and the tool's DRBW ratio sample population distributions [113,141].

PHTS's accepted allograft DRBW ratio mean was statistically smaller than what surgeons took using the tool (*C3*; ratio mean of 1.37 vs. 1.94; p-value < 0.0001; effect size = 1.20). The 1.20 effect size suggested there was a 37.8% overlap between PHTS's accepted and the tool's DRBW ratio sample population distributions [113,141]. PHTS does not collect upper DRBW ratio listings and therefore could not be compared to the tool.

UNOS's accepted allograft DRBW ratio mean was statistically smaller than what surgeons took using the tool (*C4*; ratio mean of 1.33 vs. 1.94; p-value < 0.0001; effect size = 1.39). The 1.39 effect size suggested there was between a 34.7% and 31.9% overlap between UNOS's accepted and the tool's DRBW ratio sample population distributions [113,141]. UNOS's upper DRBW ratio listing mean was slightly larger than what surgeons took using the tool but it was not statistically-significant (*C5*; ratio mean of 2.02 vs. 1.94; p-value < 0.4033; effect size = 0.13). The 0.13 effect size suggested there was between a 92.3% and 85.3% overlap between UNOS's upper listed and the tool's DRBW ratio sample population distributions [113,141].

The frequency demographics of the PCH, PHTS, UNOS accepting undersized allografts (DRBW ratio < 1.0), equal sized allografts (DRBW ratio = 1.0), and oversized allografts (DRBW ratio > 1.0) were presented in Table 6.4.

PHTS, UNOS, and Virtual Maximum DRBW Ratios

Group 1	**Virtual vs.:**					
Group 2	**PCH**		**PHTS**		**UNOS**	
	Accepted	*Upper*	*Accepted*	*Upper*	*Accepted*	*Upper*
Comparison	*C1*	*C2*	*C3*	*NA*	*C4*	*C5*
Mean 1	1.94	1.94	1.94	NA	1.94	1.94
Mean 2	1.68	2.90	1.37	NA	1.33	2.02
St. Dev. 1	0.67	0.67	0.67	NA	0.67	0.67
St. Dev. 2	0.68	1.11	0.47	NA	0.43	0.64
Sample Size 1	44	44	44	NA	44	44
Sample Size 2	44	45	1913	NA	2332	3008
St. Dev. Pooled	0.68	0.92	0.48	NA	0.44	0.64
Effect Size:	0.39	1.05	1.20	NA	1.39	0.13
P-Value:	0.0725	< 0.0001	< 0.0001	NA	< 0.0001	0.4033

Table 6.1: A series of Student T-test compared the PCH and national level DRBW ratios to the maximum TCV surgeons were willing to take during the simulated fit assessments. The *C2*, *C3*, and *C4* differences were found to be statistically-significant with effect sizes > 1 while sample sizes were N = 44. *C1* was nearly found to be statistically-significant as it nearly met the classical p-value ≤ 0.05 criterion. *C5* had a relatively low effect size and therefore was not found to be statistically-significant as was demonstrated by its p-value. *C5* was particularly interesting because the mean difference between UNOS and the tool's upper listing means was < 0.1 – the magnitude of this difference was perceived to have no clinical practical importance.

Phoenix Children's Hospital and Virtual Maximum DRBW Ratios

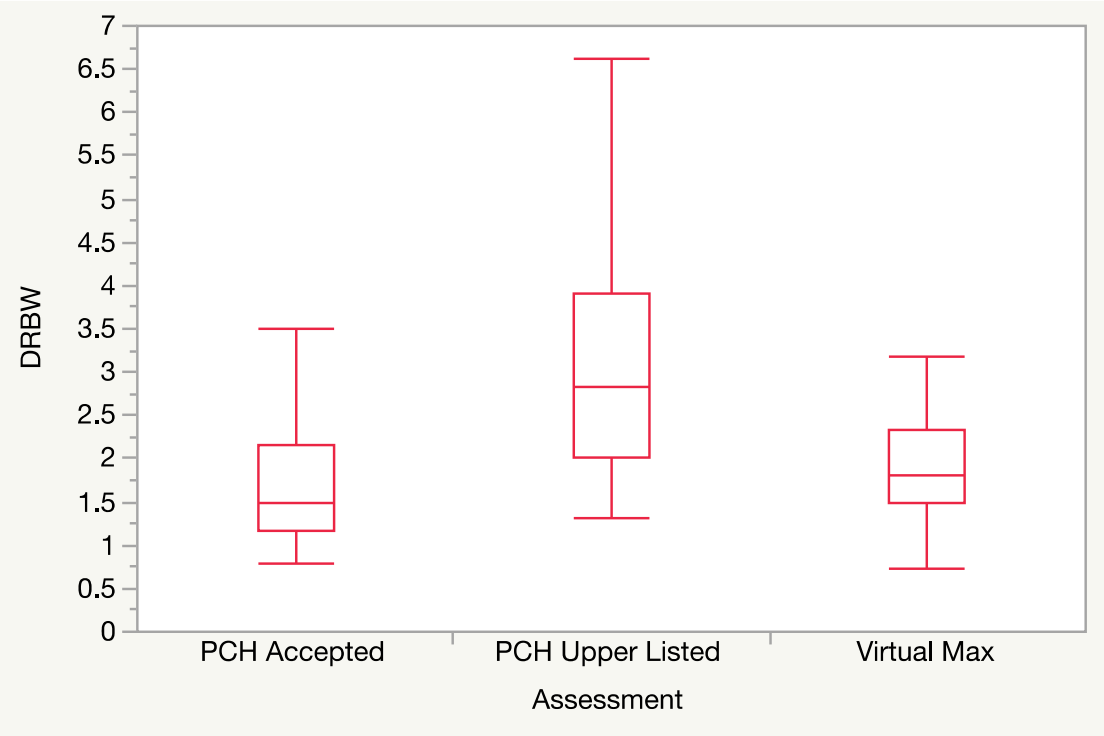

Figure 6.1: Box-plots of the PCH and virtual maximum DRBW ratios illustrated PCH's accepted and upper listed DRBW ratio population distributions were either similar or larger than the tool's DRBW ratio population distribution, in general. The results indicated the tool would not increase PCH's upper listing DRBW ratio distribution; however, pairwise comparisons were needed to determine if the tool could have expanded individual listing pools.

PCH and Virtual Maximum DRBW Ratios

	PCH		Virtual
	Accepted	*Upper*	Maximum
Maximum	3.50	6.62	3.65
3rd-Qt	2.15	3.90	2.34
2nd-Qt	1.48	2.82	1.81
Mean	1.68	2.90	1.94
1st-Qt	1.16	2.00	1.50
Minimum	0.79	1.31	0.74

Table 6.2: Figure 6.1's quintile results were presented quantitatively. In general, the tool's DRBW ratio quantiles were larger and smaller than PCH's accepted and upper listed ratio populations, respectively.

PHTS, UNOS, and Virtual Maximum DRBW Ratios

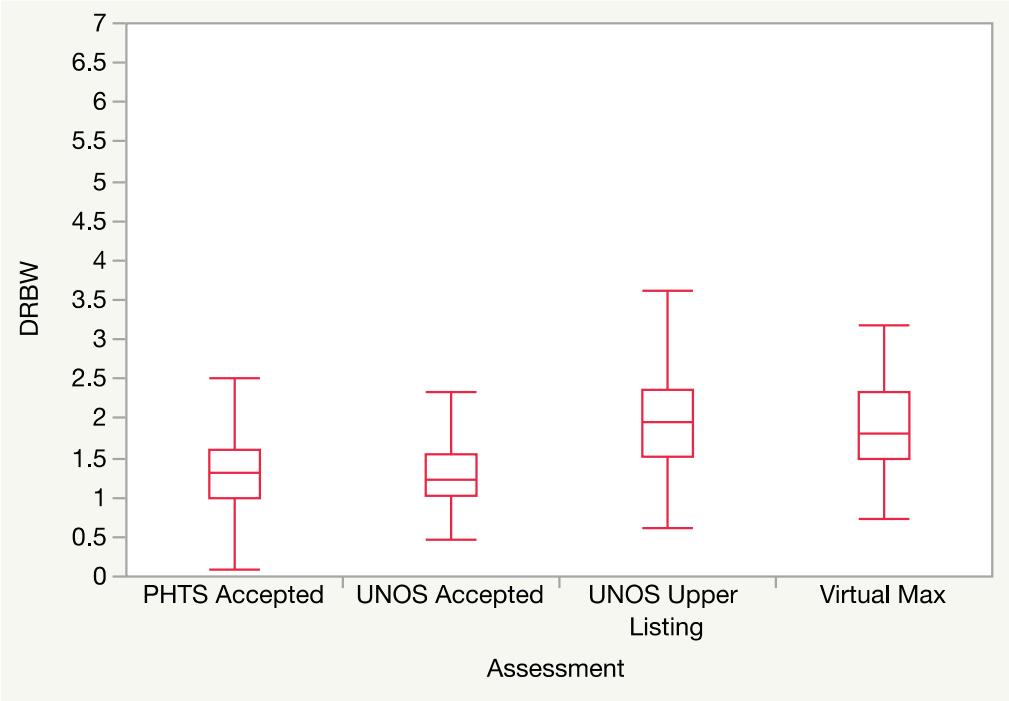

Figure 6.2: Box-plots of the national level and virtual maximum DRBW ratios. The plot illustrated the tool's DRBW ratio population distribution was generally similar to the UNOS upper listed DRBW ratio population distribution. The plot further illustrated the tool's DRBW ratio population distribution was generally larger than the PHTS and UNOS accepted DRBW ratio population distributions.

PHTS, UNOS, and Virtual Maximum DRBW Ratios

	PHTS		UNOS		Virtual
	Accepted	*Upper*	*Accepted*	*Upper*	Maximum
Maximum	3.70	NA	3.32	13.02	3.65
3rd-Qt	1.60	NA	1.54	2.37	2.34
2nd-Qt	1.30	NA	1.23	1.96	1.81
Mean	1.37	NA	1.33	2.02	1.94
1st-Qt	1.00	NA	1.01	1.53	1.50
Minimum	0.10	NA	0.48	0.60	0.74

Table 6.3: Figure 6.2's quintile results were presented quantitatively. The tool's DRBW ratio population distribution was larger than PHTS's accepted DRBW ratio population distribution. The tool's DRBW ratio population distribution was larger and slightly smaller than UNOS's accepted and upper listed ratio population distributions, respectively.

Accepted Allograft Size-Match Percentage by Classification

DRBW ratio	PCH	PHTS	UNOS
> 1.0	95.5% (42)	78.0% (1493)	75.9% (1771)
= 1.0	0.0% (0)	1.2% (22)	0.9% (22)
< 1.0	4.5% (2)	20.8% (398)	23.1% (539)
Sample Size	*44*	*1913*	*2332*

Table 6.4: A percentage (frequency) breakdown of accepted oversized, equal sized, and undersized allografts for the PCH, PHTS, and UNOS pediatric populations were presented. The acceptance of oversized allografts was constantly the norm for pediatric transplants at both PCH and the national levels, i.e., oversized allografts accounted for > 75% of pediatric transplants in each of the populations.

An investigation into the 5 comparison's effect sizes and sample sizes was performed to better understand the p-value results. Equation B.3 (see appendix B) was used to estimate the needed sample sizes to detect the mean differences between the populations with statistical-significance – this assumes the mean and variance estimates hold as the population size changes. The relationship between the needed sample size and mean difference to be detected (given the variances hold) were combined and presented in Table 6.5. To detect a DRBW ratio mean difference of 0.1 it was estimated *C4* would require the fewest individuals with approximately 411 individuals per population. Furthermore, it was approximated 1735 individuals per population would be required for a 0.1 difference to be detected in all 5 comparisons. A DRBW ratio mean difference of 0.1 was investigated because it is arguably the smallest difference of clinical interest, i.e., of practical importance. The virtual tool's sample size (N = 44) falls extremely short of these population size approximations. Using 44 individuals per population – at a minimum – the analysis suggested the smallest DRBW ratio differences the 5 comparisons could detect (based on their estimated population variances) ranged between 0.31 (*C4*) and 0.66 (*C2*). Closer inspection of the statistically-significant p-values in Table 6.1 and the smallest detectable DRBW ratio differences (when N = 44) in Table 6.5 demonstrated *C2*, *C3*, and *C4* were found to be significant because the differences they could detect were less than the mean differences between the sample populations. Although increasing sample size allows for smaller size differences to be detected it is important to recognize it does not necessarily change the populations' variances or means (and in theory it

should not). Given the variances and/or means estimates should not change then the percentage that the populations' distributions overlap should not change (see appendix B discussion) – only the difference needed to detect something with "statistical-significance" would change.

Approximate Sample Sizes Needed to Detect DRBW Ratio Differences

Group 1	**Virtual vs.:**					
Group 2	**PCH**		**PHTS**		**UNOS**	
	Accepted	*Upper*	*Accepted*	*Upper*	*Accepted*	*Upper*
Comparison	*C1*	*C2*	*C3*	*NA*	*C4*	*C5*
Mean 1	1.94	1.94	1.94	NA	1.94	1.94
Mean 2	1.68	2.90	1.37	NA	1.33	2.02
|Mean Δ|	0.26	0.96	0.57	NA	0.61	0.08
Sample Size 1	44	44	44	NA	44	44
Sample Size 2	44	45	1913	NA	2332	3008
St. Dev. Pooled	0.68	0.92	0.48	NA	0.44	0.64
	The Estimated Minimal Sample Size Per Population Group Needed to Detect a |DRBW Δ| ≥:					
0.1:	962	1735	488	NA	411	854
0.3:	113	204	58	NA	49	101
0.5:	42	75	22	NA	18	37
|Mean Δ|	149	21	17	NA	13	1321
	The Estimated Minimal Detectable |DRBW Δ| Given A Sample Size of N:					
N = 44:	0.49	*0.66*	0.34	NA	0.31	*0.46*

Table 6.5: An investigation into the relationship between sample size (per population) and the smallest DRBW ratio differences that can be detected were presented in three main blocks (separated by rows). The first block was a compilation of the key data needed for the investigation of all 5 comparisons. The second block investigated the sample sizes needed to detect various DRBW ratio differences. The row investigating the |*Mean* Δ| differences was investigating the sample size needed to detect the measured mean difference between the DRBW ratio populations. The third block investigated the smallest DRBW ratio difference that could be detected given the small population for all 5 comparisons (N = 44) and the previously estimated variances.

The PCH and national level DRBW ratios' localized averages by Wt were visualized in Figure 6.3. In general, the localized DRBW ratio averages for the accepted HTxs were less than the maximum DRBW ratio averages the tool suggest clinicians could take. The localized PCH upper listings DRBW ratio averages were larger than the maximum DRBW ratio average the tool suggest clinicians could take.

A pairwise comparison of the maximum DRBW ratios surgeons were willing to take in the virtual fit assessments suggested that 9 of the 44 PCH upper listings, i.e., > 20%, could have accepted larger donors then what they were listed for as an upper range. To have a true pairwise comparison, 1 of the 45 patients did not have images to perform a virtual fit assessment and therefore they were removed from the upper listed DRBW ratio population – hence the N = 44. The removed case was a heavily weighted outlier (DRBW ratio: 6.6) and therefore resulted in the previously presented PCH *C2* DRBW ratio mean to drop from 2.900 to 2.820. PCH's upper listing DRBW ratios for the patients the tool suggested their donor pools could have been expanded were then replaced with the virtual maximum DRBW ratios surgeons took. Replacing the PCH upper listings for the patients the tool suggested their donor pools could have been expanded to the tool's suggested maximum DRBW ratio increased the mean from 2.820 to 2.899, however, this was not statistically-significant (p-value = 0.6991, Cohen's distance effect size = 0.0827). Comparing the PCH upper listed and the tool's maximum DRBW ratios for only the 9 patients that the tool suggested their donor pools could be expanded were presented in Figure 6.4 and had an upper listed DRBW ratio mean increase from 2.08 to 2.46 but it was not statistically-significant (p-value = 0.2204, Cohen's distance effect size = 0.6012). The moderately sized effect size, i.e., 0.6012, and failure to detect a difference with statistical-significance indicated the sample size, i.e., 9 per group, was too small – it was estimated a sample size of 62 per group (124 in total) was needed to detect a difference with significance (see appendix B).

An important observation in this study was PCH's upper DRBW ratio listings were statistically larger then UNOS's reported upper listings, i.e., p-value < 0.001. The Cohen's distance effect size between PCH's upper DRBW ratio listings and UNOS's reported upper listings was 1.1232, respectively. There were 2727 (out of 3008; > 90%) cases in which UNOS had DRBW ratio

listings smaller than the PCH 2.899 mean, i.e., the PCH mean after the upper listings for the patients the tool suggested could have their donor pools expanded were replaced with the tool's suggested maximum DRBW ratio.

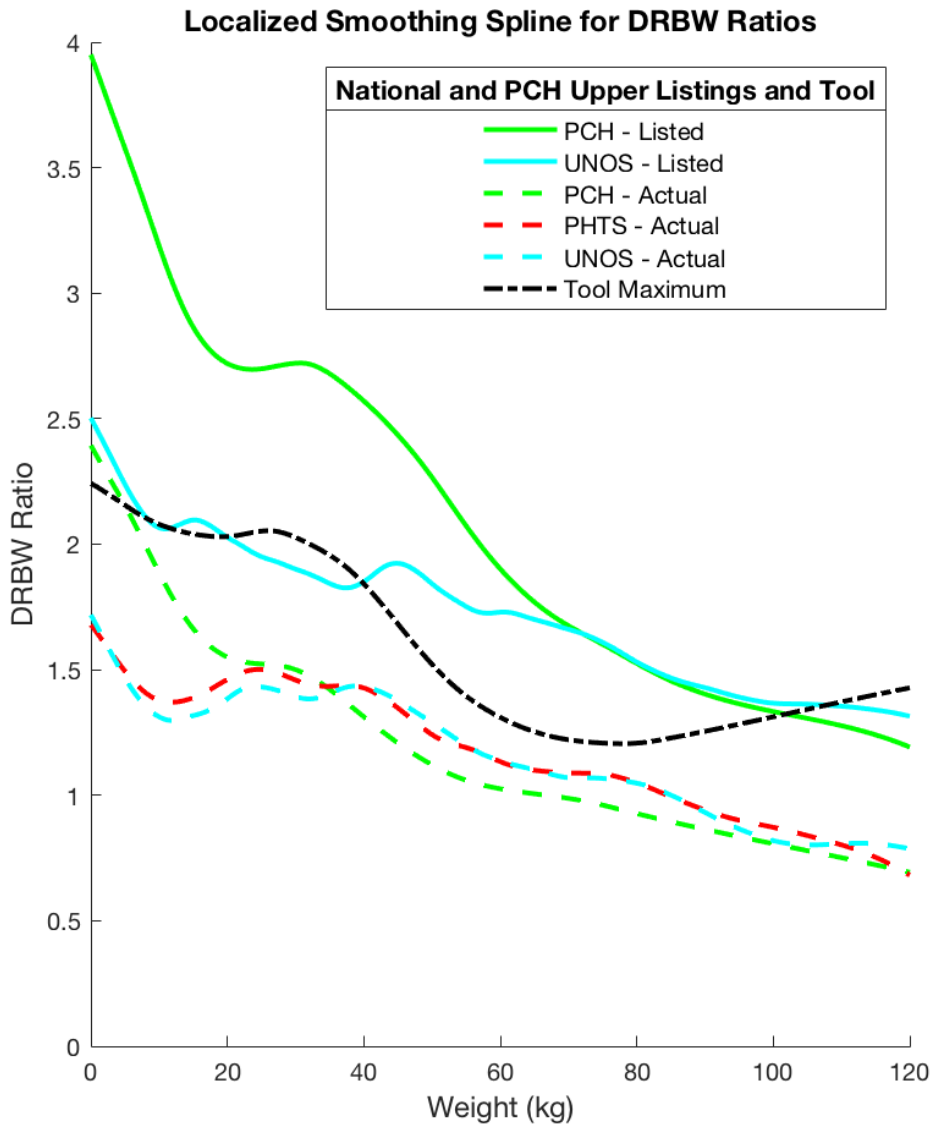

Figure 6.3: Localized spline smoothing DRBW ratio averages plotted by recipient Wt. In general, the PCH and UNOS listed localized DRBW ratio averages were equal or larger than the maximum hearts the tool identified on a localized average. In general, the actual HTx localized DRBW ratio averages were less than the maximum hearts the tool identified on a localized average.

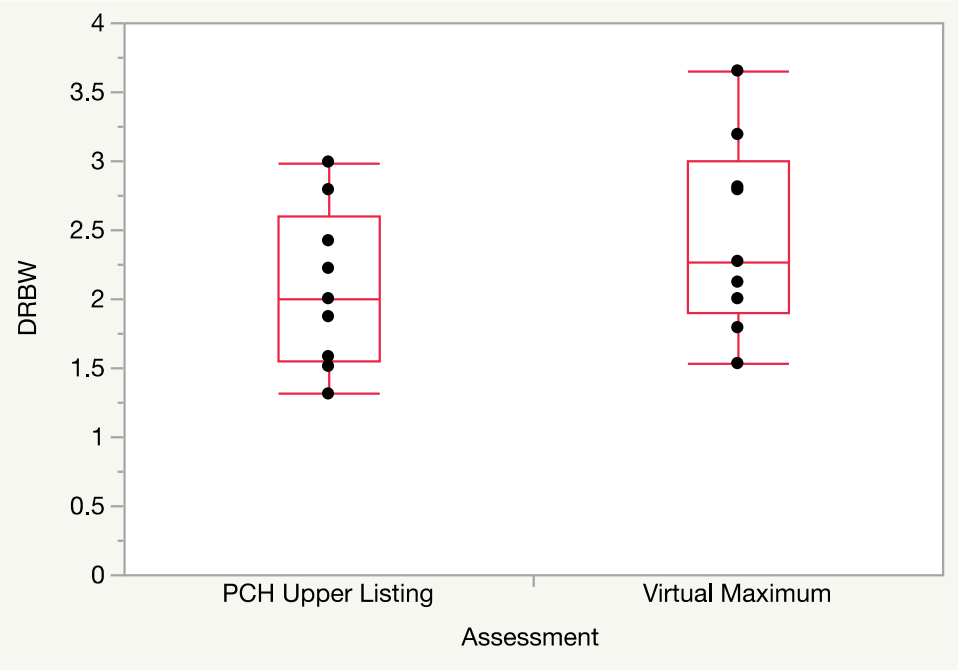

Figure 6.4: Box plot of pairwise comparison for the 9 cases (out of 44) the tool suggested PCH's upper listed DRBW ratio could be expanded. The virtual maximum DRBW ratio values were the maximum allografts surgeons were willing when using the tool.

6.5 Results for Study 2: Perceiving Fit-Related Complications

Actual donor predictors were inputted into Model B and surgeons assessed the analogue allograft's virtual fit. Out of 45 cases, 43 cases were virtually transplanted and had their outcomes analyzed, i.e., 2 cases were excluded due to unavailable data. The list of perceived fit-related complication types surgeons perceived during the virtual fit assessments and their frequencies were presented in Table 6.6. Table 6.7 compared the frequencies between actual delayed closure and any perceived fit-related complication. Delayed sternal closures were analyzed because they are a measureable outcome that can be contributed to oversized allograft complications. Delayed sternal closure and surgeon perceived complication variables had a statistically-significant association ($\mathcal{X}^2 = 18.6484$; degrees of freedom = 1; p-value = 0.000016). Of the 11 cases surgeons had perceived fit-related complications, 9 cases had delayed closures ($\mathcal{X}^2 = 4.4545$; degrees of freedom = 1; p-value = 0.034809) – this was statistically-significant. Of the 32 cases surgeons had

perceived no fit-related complications, 28 cases had no delayed closures ($\mathcal{X}^2 = 18.0000$; degrees of freedom = 1; p-value = 0.000022) – this was statistically-significant. An $\alpha = 0.05$ was the criterion used for the chi square test results to be declared as statistical-significant.

There were 4 delayed closure cases (see Table 6.7) in which surgeons did not perceive fit-related completions. Case 1 had low blood pressure and excessive bleeding. Case 2 had some ventricular dysfunction and concern for potential bleeding. Case 3 had concern for potential bleeding. Case 4 had pre-operative history of sever pulmonary hypertension. Interestingly, 2 of the 4 cases, i.e., cases 3 and 4, that surgeons did not perceive fit-related complications had a delay in sternal closure only as a precaution – not due to complications that manifested.

Frequency and Type of Surgeon Perceived Fit-Related Complications

Concern Type	Description of Surgeon's Perceived Fit-Related Compression Concern(s)	Frequency
A	Anterior-Posterior, Sternum, and Rib	4
B	Lung(s) with allograft overlap of main bronchi	1
C	Pulmonary, Pulmonary Vein	4
A and B	Concerns for A and B	0
A and C	Concerns for A and C	2
B and C	Concerns for B and C	0
A, B, and C	Concerns for A, B, and C	0

Table 6.6: Out of 43 patients there were concerns for fit-related complications in 11 patients. 9 of the 11 patients had delayed sternal closures while 2 did not have delayed closure. Chest wall (i.e., concern A) and respiratory related (i.e., concerns B and C) concerns were the 2 fundamental concern types identified by surgeons in this simulated exercise. There were 2 additional cases of perceived lung compression, however, the compressions were not perceived to affect the main bronchi (i.e., the fused allograft did not appear to overlap the main bronchi structures). Lung compression with the lack of bronchi compression to both the left and right branches was not considered a fit-related complication, e.g., it was not perceived to cause a delayed sternal closure. The 1 lung case with main bronchi compression – listed in the table – had allograft overlap on both the left and right main bronchi branches and therefore was perceived to cause delayed closure.

Delayed Sternal Closure vs. Surgeon Perceived Fit-Related Complication Matrix

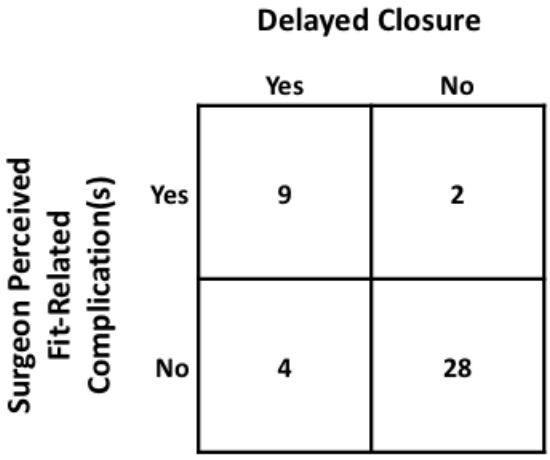

Table 6.7: There were 9 cases in which surgeons perceived fit-related complications in which the patient had delayed sternal closure.

6.6 Discussion of Tool's Clinical Utility

Study 1 and 2 results were designed to assess if the tool could be used to help expand patient donor pools safely. The results demonstrated the tool could (1) expand patient donor pools and (2) there was a significant correlation between perceived fit-related complications and delayed sternal closures.

Starting with study 1, the PHTS (*C3*) and UNOS (*C4*) accepted allograft population DRBW ratio means, in Table 6.1, were statistically smaller than the upper DRBW ratio maximum the tool suggested could be transplanted. Furthermore, the PCH (*C1*) accepted allograft p-value nearly met the $\alpha = 0.05$ criterion to state the PCH accepted DRBW ratio mean was statistically smaller than the upper DRBW ratio maximum the tool suggested could be transplanted. The tool's suggested maximum DRBW ratio mean being statistically larger, in general, than the accepted HTx means suggested clinicians were not constantly accepting the maximum allograft they could take for their patient as suggested by the tool. Figure 6.3 further demonstrated the localized (moving average) DRBW ratio means by Wt for the accepted allografts were generally less than what the tool suggested. Although accepting allografts smaller than what the tool suggested was the upper safe

limit was based on what allografts were available, the results could include cases clinicians unnecessarily passed up offers due to fear of allograft size. Future work would need to know what allograft offers were passed up to determine if clinicians were passing on allografts the tool suggested would fit during offer.

The UNOS (*C5*) upper listed allograft population DRBW ratio mean, in Table 6.1, was not found to be statistically smaller than the upper maximum the tool suggested could be transplanted. The UNOS upper DRBW ratio and the tool's DRBW ratio mean difference was < 0.1, suggesting there would be no clinically practical difference between the ratios – even if a p-value < 0.05 was found.

The PCH (*C2*) upper listed allograft population DRBW ratio mean was statistically larger than the upper maximum the tool suggested could be transplanted. Interestingly, > 95% of PCH's accepted allografts were oversized (see Table 6.4) with an accepted DRBW ratio mean of 1.68 even though PCH's upper listed DRBW ratio mean was 2.90. Furthermore, PCH's accepted DRBW ratio mean was between the PHTS and UNOS accepted DRBW ratio means and the tool's suggested upper DRBW ratio mean. Future work would need to know what allograft offers were passed up but these current findings could be explained with the new hypothesis that even though PCH list high they might be passing on allografts during offer that the tool would have suggested would fit.

A pairwise comparison of maximum DRBW ratios surgeons were willing to take in the virtual fit assessments suggested > 20% of PCH's listed patients, i.e., 9 of 44, could have accepted larger hearts than what they were listed for. The result of this pairwise comparison suggested there was a subset of the PCH population that the tool could have expanded their listed donor pools. It was an interesting find that there was a subset in PCH's listed population that the tool suggested their donor pool listing range could have been expanded even though this institute is aware they are aggressive in their listing. PCH's originally listed DRBW ratio mean difference, when compared to the national level, was > 0.85 with statistical-significance – this supported the fact that PCH is aggressive in its upper DRBW ratio listing. In fact, results showed > 90% of the patients in UNOS database had listed DRBW ratios less than the average of the tool's listed DRBW ratio for PCH

(i.e., the 2.899 listed DRBW ratio average after replacing the PCH listed DRBW ratios with the maximum virtual DRBW ratio for the cases the tool suggested PCH listings could be expanded). Given:

1) the tool would expand a subset of the PCH listing population,

2) the PCH listing population DRBW ratio mean was statistically larger than that of the national level, and

3) > 90% of the cases in the national database population had listed DRBW ratios less than PCH's listed DRBW ratio when the tool was used,

the results might imply the tool would have potential to expand national donor pools; however, future work needs to investigate this.

In analyzing the p-value results of the 5 comparisons in Table 6.1, it was suggested that increasing the sample size would allow for population mean differences to be found statistically-significant. Sample size controlling if a mean difference was found to be statistically-significant assumed the population means and variances were held constant – this was further covered in appendix B. However, it is important to note that increasing sample population sizes, in theory, does not generally change the populations' means or reduce the populations' variances and therefore would not change the overlap of the two distributions. In other words, increasing the sample size would not improve one's chances, i.e., probability, of correctly guessing which population a new measurement belongs to even though one increases the chances that a p-value < 0.05 will be calculated. The consequence is increasing the sample size increases the chances the mean difference of 2 or more populations is found to be of statistical-significance.

Study 2 virtual fit assessments results in Table 6.6 demonstrated chest wall, lung (with allograft overlap of main bronchi), and pulmonary compression were the types of fit-related

complication concerns surgeons perceived during the virtual HTx fit assessment. Although a wide range of post-operative complications (e.g., mechanical support and pulmonary vein stenosis) were originally considered for indications of allograft oversized complications, it was advised the study should focus on delayed sternal closures. Many of the original post-operative complications considered as metrics for fit-related complications (including delayed sternal closure) could be caused by non-fit-related complications. However, delayed sternal closure was perceived to be the measure that most likely corresponded to oversized allograft fit-related complications (relative to other metrics) and therefore was the only outcome metric analyzed herein. Furthermore, there were 2 cases of perceived lung compression but the fused allograft did not appear to overlap the main bronchi – these cases were not perceived to cause delayed chest closures and therefore they were not included in Table 6.6. Results in Table 6.7 demonstrated there was a statistically-significant association between the perceived fit-related complications and delayed sternal closure variables, i.e., p-value = 0.000016. Of the cases surgeons perceived there would be fit-related complications, the frequency of delayed sternal closures was statistically-significant, i.e., p-value = 0.034809. Of the cases surgeons perceived there would be no fit-related complications, the frequency of not having a delayed sternal closure was statistically-significant, i.e., p-value = 0.000022.

Out of the 9 cases that (1) surgeons perceived fit-related complications and (2) had outcomes with delayed closure, as shown in Table 6.7, only 1 of these closures had actual fit-related complications (or indications of fit-related complications) reported in the surgical notes. Bleeding and hemodynamic concerns were the causes for the other 8 cases to have delayed closures. Although, there were no clinical indications available to suggest the bleeding or hemodynamic issues for the 8 cases were related to allograft oversizing, this might be a multi-factorial problem. This might be a multi-factorial problem because there might be unknown nuisance factors, related to oversized allograft complications, that caused the bleeding and hemodynamic issues. What was observed when surgeons perceived or did not perceive fit-related complications was their perception corresponded with statistical-significance to patients having or not having delayed closures, respectively. Future work would need to identify what, if any,

oversized allograft nuisance factors related the bleeding and hemodynamic complication trends to delayed sternal closure.

6.7 Discussion on Virtual Transplant Case Study Using Donor Images and the Tool

The virtual HTx fit assessment tool was developed to utilize a healthy heart library and an allograft TCV prediction model developed in chapters 4 and 5, respectively. The need to identify a healthy heart reconstruction (of similar TCV) to serve as an allograft analogue was to overcome the technical challenges in performing a virtual fit assessment with donor images. The technical challenges include having donor images readily available for a fit assessment. The assessment would need to be performed within the typical 1-hour time-window clinicians have available to make a provisional acceptance or refusal of an offer. Furthermore, once a donor is deceased, i.e., available for offer, the financial cost of sustaining the donor for procurement falls to the donor's center. The current system does not financially support donor centers to acquire CT or MR images for a fit assessment as this is an atypical request. Nevertheless, there was one unique case in which donor images became available and used in supplemental patient care (previously presented as a case study [142]).

The case was of a hypoplastic left heart patient (female, 10 years old, 140cm, 28kg) that was offered a larger donor (female, 16 years old, 163cm, 60kg). Given the complex anatomical nature of the hypoplastic left heart patient's cardiothoracic space, the clinical team was leaning towards declining the donor offer due to a 2.1 DRBW ratio. A virtual HTx fit assessment using donor images, illustrated in Figure 6.5, and the corresponding donor-recipient TCV ratio, i.e., 0.6, were used during supplemental patient care. The additional information augmented the clinical team's available information. The clinical team accepted the allograft offer with no observed post-operative fit-related complications. The patient was discharged 15 days post-transplant and is currently thriving well over a year. To achieve this case's virtual assessment, a provisional yes was made such that time was allowed for the electronic data transfer of the donor images and virtual assessment to be performed before a final allograft acceptance was made.

Virtual Fit Assessment for Case Study

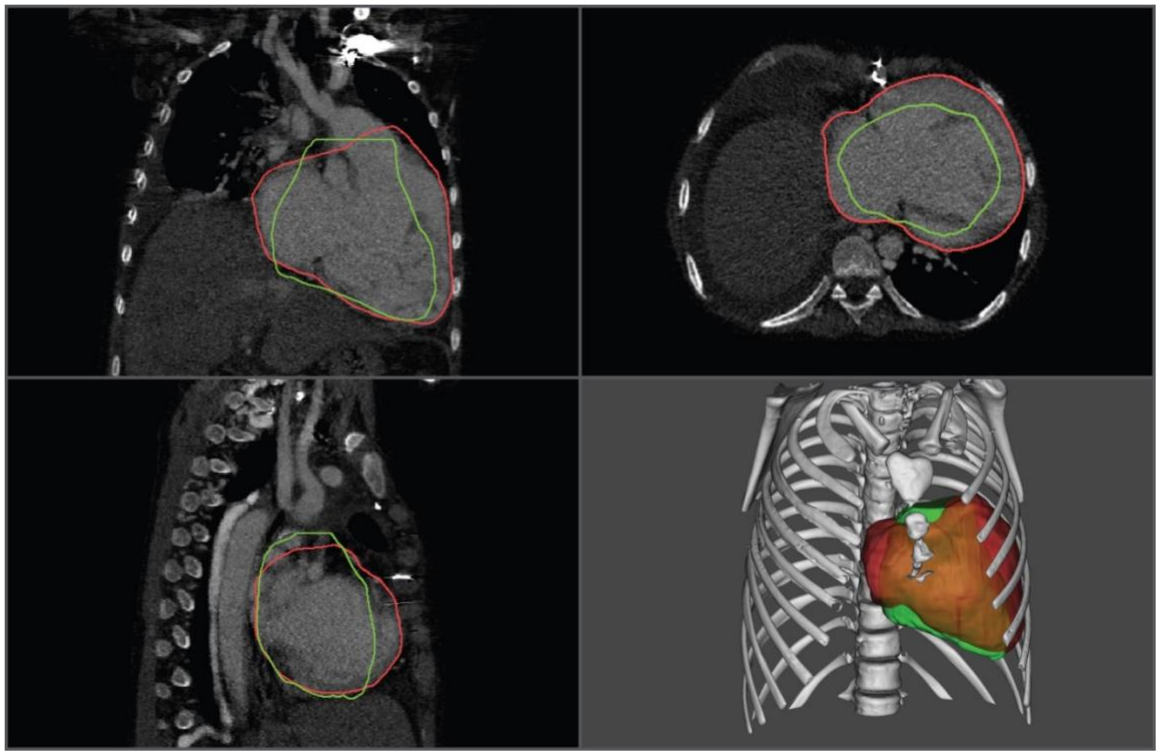

Figure 6.5: The case study virtual HTx fit assessment in which the allograft (green) was reconstructed from donor images. The allograft is fused onto the recipient's pre-transplant CT images with the native heart (red). The allograft was found to be undersized and therefore augmented the clinical team's perception to accept the offer. The image was previously presented but did not require permissions to reproduce (see appendix D) [142].

Retrospectively inputting donor's predictors into Model B resulted in a -45% error, i.e., over-prediction. This large over-prediction error was unsurprising because the largest over-prediction error measured in Model B's training dataset was -56%. These large errors are still present even though Model B had training MAPE and sMAPE values of 12.7% and 12.3%, respectively. Interestingly, the large over-prediction of Model B still resulted in a 0.85 donor-recipient TCV ratio and therefore suggested the allograft was still undersized. Even if the allograft over-prediction did not deem the transplant permissive, declining an offer due to excessively oversized concerns is preferable to accepting an allograft that was larger than expected.

There are two key messages from this unique case study. First, the case study suggested there might be scenarios in which virtual HTx fit assessments can help clinicians perceive acceptable fits that they would otherwise not take. In this case the clinical team was originally

concerned for fit-related complications (possibly including delayed closure) based on (1) pathology and (2) the DRBW ratio. However, the reconstructed allograft was very undersized and therefore the virtual fit assessment did not suggest the offered allograft would have trouble fitting the native heart's space. It is important to point out that this case study did not perceive fit-related complications and therefore does not indicate if the tool could perceive fit-related complications. The case study did support study 1 and 2's findings that the tool can be used to identify a subset of cases in which the donor pool can be safely expanded. Second, although using a prediction model might be clinically preferable to the alterative, it is a model that will not work perfectly in every scenario. A push to acquire donor images for virtual HTx fit assessments might lead to better patient outcomes as clinicians might be able to better perceive fit-related complications before transplant.

CHAPTER 7

LIMITATIONS, FUTURE WORK, AND CONCLUSIONS

There were two hypotheses laid out in chapter 1 for this thesis. The first hypothesis was a tool leveraging medical images could be developed to allow clinicians to virtually assess donor allograft fits with qualitative information. The second hypothesis was the tool would help clinicians expand patient donor pools and perceive fit-related complications. It was found the tool (1) could have expanded the upper DRBW ratio listing range for > 20% of PCH's patients in the study and (2) a statistically-significant trend between delay sternal closures and surgeon perceived fit-related complications (using the tool) was observed. The results demonstrated a novel tool could be developed to leverage medical images and help clinicians assess allograft fits, i.e., the first hypothesis was achieved. The results further supported the tool could be used to expand patient donor pools, i.e., the first part of the second hypothesis was achieved. Lastly, the results found a trend between surgeon perceived fit-related complications and delayed sternal closure, however, future work needs to investigate this multi-factorial relationship to confirm causality, i.e., the second part of the second hypothesis looks promising but future work is needed to fully answer this question.

Chapter 7 covers the limitations, future work, and final conclusions of the author's thesis work herein. The limitations will cover the healthy heart library and its possible effects on allograft reconstructions and the allograft TCV prediction model. Furthermore, the limitations section will cover clinical limitations and their possible effects on the analysis of the tool's clinical utility. Next, future work will look at (1) improving the clinical utility study, (2) improving the novel virtual fit assessment tool for HTx and other areas of solid-organ transplant, and (3) investigate new areas/questions that arose from this work. The chapter and therefore the thesis will conclude with a discussion on what has been achieved from the engineering and/or scientific points-of-view.

7.1 Limitations of the Healthy Heart Library and Allograft TCV Prediction Model

The healthy heart library and the allograft TCV prediction model relied on 97 “healthy” heart patients identified to develop the novel virtual HTx fit assessment tool. The limited sample size and source of population were key limitations in this body of work. The type of available data for a retrospective study further limited the modeling process.

For logistical reasons, the work herein was designed for a retrospective, single center even though it was recognized there would be a limited number of “healthy” heart patients available in a chart review. K-fold cross-validation and parametric modeling techniques were intentionally deployed to develop and validate an allograft TCV prediction model with a limited data sample size. The 97 patients identified were above the 50 minimum typically suggested by statisticians when building a linear, parametric regression model [105]. Although steps were taken to minimize the limitations of working with a small but realistic sample size, future work might want to further expand upon the size of the healthy heart library in a multicenter study. Increasing the library size would allow for more complex modeling techniques to be implemented.

The inclusion criteria for “healthy” was another limitation of the healthy heart library. In this retrospective study, patients with CT or MR images were included into the library. Given these subjects were patients, it indicates (1) they are not truly healthy or (2) there was a concern they may not be healthy. Careful chart evaluation from radiology interpretation summaries were reviewed to help ensure included individuals had normal, well-functioning, healthy hearts with normal anatomical volumes pertaining to the chambers and overall TCV. Pathologies and/or treatments known to affect heart anatomy and/or physiology were automatically excluded. Mild aberrations of the greater vessels and coronaries were generally considered healthy unless indicated otherwise. The careful chart review helped prevent the inclusion of unhealthy hearts, however, future work may want to develop a prospective study to acquire truly heart healthy individuals.

The retrospective study and the tool were limited on what data was available to be used as predictors in the allograft TCV prediction model. The limitations arose from both the chart review

and what data is typically available from donors during offer. For example, donor echocardiograph data is available at offer and includes the acquisition of a measured left ventricular volume. However, it was known during study development that this ventricular measurement would not be available in the majority of subjects in the healthy heart library as they were not necessarily likely to have received an ultrasound of the heart. The development of the allograft TCV prediction model was carefully crafted based on what donor parameters would be available from subjects in the healthy heart library and at donor offer. Unfortunately, due to the lack of echocardiograph and other cardiac based measurements, the model was not developed from direct cardiac measurements but rather relied on secondary, gross body metric parameters, e.g., Sex, Ht, Wt, etc. Future work might want to consider the prospective development of a healthy heart library in which subject-specific cardiac parameters can be included into the model develop - assuming the corresponding donor measurements are readily available as standard of care during offer.

The library and model limitations were carefully considered during the development of the novel tool. The author made careful considerations, based on the data that would be available for this initial, retrospective project, to develop a novel virtual HTx fit assessment tool. Considerations included the fact that donor images may or may not be available at the time of donor offer for a virtual fit assessment. The author took several additional steps (e.g., careful healthy heart inclusion criteria, statistical model design, validation steps, etc.) to help ensure a sound, practical allograft TCV prediction model was developed (e.g., as the 12.3% sMAPE suggested) given the current limitations.

7.2 Limitations of the Clinical Utility Assessment

The initial evaluation of the tool's clinical utility had 2 major sources of limitation. The first limitation was the sample size of the data that was available. The second limitation was what specific type of data was available based on current clinical practices (e.g., what data gets measured and/or recorded). Both of these limitations were ultimately due to working with historical data in a retrospective study.

Sample size was largely driven by PCH's HTxs that had pre-operative CT or MR image images of the cardiothoracic space. Between 45 and 43 PCH HTx individuals were used in the various studies (and sub-study comparisons) based on what data was available. This low PCH population size (along with low effect sizes < 0.5) resulted in many of the mean differences in Table 6.1 to not be statistically-significant. However, increasing the sample size to get a statistically-significant p-value would not have changed the probability of guessing which population a new measurement belonged to, in theory. Another issue with the small PCH sample size was whether the sample means and variances truly represent PCH's population. Future work might need to either include more individuals, repeat the study with a new PCH cohort, or perform a cross-validation to help confirm the true PCH means and variances within this study.

The listed, accepted, and virtual maximum DRBW ratios represented fundamentally different scenarios and therefore were not completely comparable. The listed DRBW ratio corresponded to a clinical scenario in which clinicians were preparing for a "last-ditch effort" wherein the probably of declining an offer could have resulted in a poorer outcome. However, if the patient isn't sick enough to accept a "last-ditch effort" allograft then decision makers will often pass on offers approaching the upper DRBW ratio limit. Additionally, in conversations with a PCH clinician, it was explained that actively over-listing patients has no repercussions – other then maybe clinicians having to spend more time declining offers – and would only provide the institute with more opportunities to procure allografts for their patients. The upper listed DRBW ratio was further complicated because clinicians may go back and increase the upper listing range depending on how sick the patient has become. The accepted DRBW ratio represents allografts that were within the upper and lower listed ranges that were actually accepted – it states nothing on what offers were actually declined even though they were within the listing range. The "level of sickness" is one hidden factor contributing to why an allograft of given DRBW ratio gets accepted (or declined). Lastly, the virtual tool asked surgeons to push the upper allograft limits within reason, i.e., no "last-ditch effort" scenarios.

Comparing the listed and accepted DRBW ratio means to the tool's suggested DRBW ratio means might provide some insight into the tool's clinical utility, however, it would be limited. To

address the limitation of interpretation for these population mean comparisons, a pairwise comparison between PCH's upper listed and the tool's suggested DRBW ratios was performed to answer if the tool could expand patient donor pools. The pairwise comparison analyzed the frequency (and percentage) of PCH's listed population that the tool suggested could have had their listed DRBW ratio expanded.

During virtual fit assessments with the surgeons it became clear that asking them to perceive fit-related complications was not as trivial as originally thought. There was uncertainty as to whether or not the level of compression the surgeon was perceiving would result in problematic outcomes. Surgeons were still asked to make a judgment call in this study without formal training. The interactions with surgeons suggested (1) a formal, pre-clinical prospective study on what oversized allografts look like in a virtual assessment might be warranted and (2) surgeons might benefit from using the pre-clinical results to train on how to assess the virtual fits. A future, pre-clinical or clinical prospective study might find that the virtual maximum surgeons accepted in the simulated scenarios herein were not the true virtual maximum allografts that could be taken. Furthermore, the maximum allografts surgeons perceived they could take were based on perceived compression effects; it is likely some compression effects with post-operative complications are preferable to declining an offer. Clinicians will need to (1) better define what a poor outcome is and (2) better determine how to perceive these defined poor outcomes from medical images, for the full potential clinical utility of the proposed tool to be realized.

The use of measureable outcomes with known (or perceived) associations to oversized fit-related complications was a technical challenge for this study. First, many of the complications associated with oversized allografts, i.e., respiratory related complications, were clinically perceived by PCH personal to be more likely indicate something other than oversized fit-related complications. Delayed sternal closures were the only outcome analyzed herein because it was perceived (before the analysis was performed) to be the outcome most likely to correspond with oversized fit-related complications. Second, of the 9 delayed closures that the surgeons had also perceived fit-related complications (while blinded to actual outcomes), only 1 case had surgical notes indicating oversized allograft complications. The other 8 cases indicated bleeding and

hemodynamic issues for the delayed closure – the complications may or may not have been part of a multi-factorial interaction caused by the oversized allograft. This study only states the cases that surgeons had perceived fit-related complications also had delayed closures with statistical-significance – a prospective study would be needed to fully understand if the trend had causality as the retrospective surgical notes might be limited in information.

7.3 Future Work

There are 5 areas of future work that can be derived from this thesis. Areas of future work to improve the current work focus on expanding the healthy heart library pool, improving studies on the clinical utility of the tool, and applying more advanced modeling techniques to predict allograft TCV and clinically safe donor fits. The need to model pediatric, biological scaling relationships of cardiac structures from more advanced CT and MR images in future work also arose from this study

First, as pointed out in the limitation sections, increasing the healthy heart's library size and inclusion of cardiac and other closely related parameters is warranted in future work – this could be achieved in a prospective study that recruits healthy heart subjects. Increasing the library's size will allow for more advanced statistical modeling techniques to be implemented and result in a larger training dataset that would likely help improve allograft prediction performance. Cardiac (e.g., left ventricular volume from echocardiograph) and anatomical measurements near the heart (e.g., chest circumference) could be acquired to help further predict TCV. The inclusion of echocardiograph measurements was originally thought of, but the perceived lack of these echocardiograph measurements for healthy heart individuals was a limitation for the retrospective study. Similarly, cardiothoracic related measurements – such as chest cavity circumference – could further help with TCV prediction but body type (from subcutaneous fat) and relative heart size (varied during development) would need to be carefully accounted for [143–145].

The developed tool herein focused on expanding donor pools based on geometric concerns, i.e., oversized allograft sizes, but did not focus on cardiac output needs. Both over- and

under-perfusion – due to inappropriate cardiac output needs – have warranted clinical concern in the literature [57,59,67]. Future work might consider inclusion of cardiac output in the virtual fit assessment process to help ensure a donor offer would provide the patient with an appropriate blood flow rate. Cardiac index might be a better metric to account for cardiac output as it is a clinically accepted metric that indexes cardiac output needs to patient size, i.e., BSA.

Second, additional assessment of the tool's clinical utility in a prospective study is warranted before it can be determined if the tool can help clinicians to safely expand patient donor pools by determining if they can actually perceive post-operative fit-related complications by using the tool. This thesis only demonstrated a statistically-significant trend between delayed closure and surgeon perceived fit-related complications using the tool – future work needs to investigate and confirm actual causality. A study into what an oversized allograft with complications looks like is warranted as perceiving safe fits does not appear to be trivial. A prospective study designed to record specific allograft HTx outcomes (including both size appropriate and mismatched) would help to determine if the tool does in fact help clinicians to perceive fit-related complications. The prospective study would need to better determine what outcomes would need to be measured and how to measure them to assess if surgeons could perceive fit-related complications. Metrics identified as indicators of poor outcome could be used to help predict poor outcomes in a support vector machine classifier and therefore determine if an offer should be accepted. The HTx performance metrics could be acquired from the virtual fit assessment or from other clinical metrics, e.g., cardiac index. Several support vector machine classifier models could be developed to determine if an offer should be accepted based on a patient's degree of sickness, e.g., A1, B1, and B2 listings. It is likely that the classifier would more likely accept mismatched allografts in sicker patients as the risk of accepting the allograft is less than if the allograft was reject.

Third, more advanced TCV prediction modeling process might be warranted in which small, clinically acceptable TCV prediction errors do not add to the cost function, i.e., only prediction errors that are unacceptably large would penalize the model. Support vector regression would allow for small errors (i.e., errors that have no importance in the practical sense) to not add to the cost function [146,147]. This regression technique allows for coefficient estimates to be approximated

in such a manner that reduces errors of practical importance and does not estimate coefficients to address errors that have no practical importance [118,147]. In other words, the method focuses on minimizing the chance of making large, problematic errors at the cost of not improving small errors that have no real practical consequence. To implement this regression technique, clinicians will need to identify an acceptable prediction error size – the size error could either be a constant or non-constant value. A benefit of support vector regression techniques is overall model robustness can potentially be improved as a consequence of the modeling process not caring about small errors that have no practical importance [146,148]. Furthermore, support vector regression models can easily model both linear and nonlinear relationships [146].

Segmented regression (also known as piecewise regression) is another modeling technique that may have benefits in TCV prediction. It is known that rapid developmental growth happens in (1) infant and early childhood and (2) during puberty. These periods of development may have different rates of growth that were not analyzed herein – in part because of the limited healthy heart library size. After a further increase in the healthy heart library's population size, it would be advisable to use segmented regression to capture the unique differences between the different periods of human growth.

Fourth, future work might want to consider if virtual fit assessments can be improved upon and/or expanded to other solid-organ transplant areas. It might be found that finite element analysis studies, using material property parameters, could help to improve fit assessments. Lung, liver, and kidney transplants are solid-organ transplant areas that could also benefit from a virtual fit assessment tool to help improve patient outcomes, particularly in the pediatric arena.

Fifth, in chapter 4 a 0.8 scaling signal was observed several times in the healthy heart library data. Future work should determine if this previously discussed 0.8 signal was happenstance or if it has biological significance. Furthermore, future work should investigate pediatric cardiac signals from CT and or MR images. The author found most reported cardiac scaling relationships were limited in pediatrics, i.e., cardiac scaling relationships were typically available from adult studies, and that such signals were often acquired using echocardiograph data. However, the

clinical utility of CT and MR data is growing in the current era while the availability of scaling relationships derived from advanced cardiac CT or MR imaging is limited.

The author proposes the immediate next sets for the further improvement and validation of the novel tool developed herein is a multicenter, prospective study. To improve TCV prediction, the prospective study should (1) recruit health individuals, e.g., no patients, and (2) include both cardiac and cardiothoracic measurements that could easily be made readily available at donor offer, e.g., from echocardiograph measurements. At the same time clinicians will need to identify, e.g., classify, poor and good outcomes based on patients' categorical levels of sickness in a large HTx sample population – patients will need to have HTx outcomes. The classification of poor and good outcomes could then be used to develop a support vector machine classifier to guide the acceptance or decline of donor offers. The author suggest the future work should consider support vector machine classification techniques and not logistics regression techniques due the wide variety of nonlinear kernels having been previously developed for the former classification method [118]. The clinical utility can then be assessed to determine if (1) the classifier still allows clinicians to accept allografts they normally would not accept based on the DRBW ratio and (2) determine the tool's sensitivity and specificity to predicting both poor and good outcomes. The next steps for both the regression model development and assessment of the tool's clinical utility in HTx assessments will need large sample sizes due to effect sizes being either of "small" or of "moderate" size, in general.

7.4 Conclusions

A novel virtual heart transplant fit assessment tool was developed in this thesis. From this work, knowledge about pediatric cardiac biological signals was gained and a model to predict TCVs during human development was developed. In addition to human development, the TCV model also considers body type, e.g., overweight.

Assessment of the tool's clinical utility suggested the novel tool would have expanded the PCH listed DRBW ratios for > 20% of the PCH patients in this study. This demonstrated the tool

could have expanded the donor pool for a subset of PCH's HTx population. Furthermore, an assessment of the tool's utility found a trend between surgeon perceived fit-related complications and delayed sternal closure, i.e., a clinical outcome perceived to be associated with oversized allograft transplants, was observed with statistical-significance. The early retrospective findings suggest the novel tool developed herein can possibly help clinicians to safely expand their patient donor pools but future, prospective work is warranted to further assess the tool's clinical utility. The DRBW ratio metric would still be a clinically important metric in the foreseeable future, however, the results suggest the virtual tool can be incorporated with the DRBW ratio metric to help expand patient donor pools. We expect that only after experience will clinicians begin to trust the tool developed herein; in the imminent future clinicians will likely use the tool conservatively in prospective research studies and supplemental patient care.

REFERENCES

[1] Lietz K and Miller L W 2007 Improved Survival of Patients With End-Stage Heart Failure Listed for Heart Transplantation *J. Am. Coll. Cardiol.* **50** 1282–90

[2] Singh T P, Almond C S, Taylor D O and Graham D A 2012 Decline in Heart Transplant Wait List Mortality in the United States Following Broader Regional Sharing of Donor Hearts *Circ. Heart Fail.* **5** 249–58

[3] Khush K K, Menza R, Nguyen J, Zaroff J G and Goldstein B A 2013 Donor Predictors of Allograft Use and Recipient Outcomes After Heart Transplantation *Circ. Heart Fail.* **6** 300–9

[4] Singh T P, Milliren C E, Almond C S and Graham D 2014 Survival Benefit From Transplantation in Patients Listed for Heart Transplantation in the United States *J. Am. Coll. Cardiol.* **63** 1169–78

[5] Ojo A O, Heinrichs D, Emond J C, McGowan J J, Guidinger M K, Delmonico F L and Metzger R A 2004 Organ donation and utilization in the USA *Am. J. Transplant.* **4** 27–37

[6] Almond C S D, Thiagarajan R R, Piercey G E, Gauvreau K, Blume E D, Bastardi H J, Fynn-Thompson F and Singh T P 2009 Waiting List Mortality Among Children Listed for Heart Transplantation in the United States *Circulation* **119** 717–27

[7] Schmidt-Nielsen K 1984 *Scaling: Why is Animal Size so Important?* (Cambridge University Press)

[8] Allen D H 2013 *How Mechanics Shaped the Modern World* (Springer Science & Business Media)

[9] Arvand A, Hormes M and Reul H 2005 A Validated Computational Fluid Dynamics Model to Estimate Hemolysis in a Rotary Blood Pump *Artif. Organs* **29** 531–40

[10] Ayalew T B, Krajewski W F and Mantilla R 2014 Connecting the power-law scaling structure of peak-discharges to spatially variable rainfall and catchment physical properties *Adv. Water Resour.* **71** 32–43

[11] Bejan A and Lorente S 2006 Constructal theory of generation of configuration in nature and engineering *J. Appl. Phys.* **100** 041301

[12] Borowy B S and Salameh Z M 1996 Methodology for optimally sizing the combination of a battery bank and PV array in a wind/PV hybrid system *IEEE Trans. Energy Convers.* **11** 367–75

[13] Caduff M, Huijbregts M A J, Althaus H-J, Koehler A and Hellweg S 2012 Wind Power Electricity: The Bigger the Turbine, The Greener the Electricity? *Environ. Sci. Technol.* **46** 4725–33

[14] Chakraborty S 2011 *Mechanics Over Micro and Nano Scales* (Springer Science & Business Media)

[15] Lepist E-I and Jusko W J 2004 Modeling and allometric scaling of s(+)-ketoprofen pharmacokinetics and pharmacodynamics: a retrospective analysis *J. Vet. Pharmacol. Ther.* **27** 211–8

[16] West G B and Brown J H 2005 The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization *J. Exp. Biol.* **208** 1575–92

[17] Paton K R, Varrla E, Backes C, Smith R J, Khan U, O'Neill A, Boland C, Lotya M, Istrate O M, King P, Higgins T, Barwich S, May P, Puczkarski P, Ahmed I, Moebius M, Pettersson H, Long E, Coelho J, O'Brien S E, McGuire E K, Sanchez B M, Duesberg G S, McEvoy N, Pennycook T J, Downing C, Crossley A, Nicolosi V and Coleman J N 2014 Scalable production of large quantities of defect-free few-layer graphene by shear exfoliation in liquids *Nat. Mater.* **13** 624–30

[18] Rakhmatov D, Vrudhula S and Wallach D A 2003 A model for battery lifetime analysis for organizing applications on a pocket computer *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **11** 1019–30

[19] Kandasamy R, Periasamy K and Sivagnana Prabhu K K 2005 Chemical reaction, heat and mass transfer on MHD flow over a vertical stretching surface with heat source and thermal stratification effects *Int. J. Heat Mass Transf.* **48** 4557–61

[20] Stahl W R 1967 Scaling of respiratory variables in mammals *J. Appl. Physiol.* **22** 453–60

[21] Stahl W R 1965 Organ Weights in Primates and Other Mammals *Science* **150** 1039–42

[22] Anderson B J and Meakin G H 2002 Scaling for size: some implications for paediatric anaesthesia dosing *Pediatr. Anesth.* **12** 205–19

[23] Bide R W, Armour S J and Yee E 2000 Allometric respiration/body mass data for animals to be used for estimates of inhalation toxicity to young adult humans *J. Appl. Toxicol.* **20** 273–90

[24] Blinman T and Cook R 2011 Allometric Prediction of Energy Expenditure in Infants and Children *ICAN Infant Child Adolesc. Nutr.* 1941406411414416

[25] Huang Q, Gehring R, Tell L A, Li M and Riviere J E 2015 Interspecies allometric meta-analysis of the comparative pharmacokinetics of 85 drugs across veterinary and laboratory animal species *J. Vet. Pharmacol. Ther.* **38** 214–26

[26] Johnson T N 2008 The problems in scaling adult drug doses to children *Arch. Dis. Child.* **93** 207–11

[27] Lindstedt S L and Schaeffer P J 2002 Use of allometry in predicting anatomical and physiological parameters of mammals *Lab. Anim.* **36** 1–19

[28] Mahmood I 2014 Dosing in Children: A Critical Review of the Pharmacokinetic Allometric Scaling and Modelling Approaches in Paediatric Drug Development and Clinical Settings *Clin. Pharmacokinet.* **53** 327–46

[29] Yassen A, Olofsen E, Kan J, Dahan A and Danhof M 2007 Animal-to-Human Extrapolation of the Pharmacokinetic and Pharmacodynamic Properties of Buprenorphine *Clin. Pharmacokinet.* **46** 433–47

[30] Elser J J, Fagan W F, Kerkhoff A J, Swenson N G and Enquist B J 2010 Biological stoichiometry of plant production: metabolism, scaling and ecological response to global change *New Phytol.* **186** 593–608

[31] West G B, Brown J H and Enquist B J 1997 A General Model for the Origin of Allometric Scaling Laws in Biology *Science* **276** 122–6

[32] Rich P R 2003 The molecular machinery of Keilin's respiratory chain *Biochem. Soc. Trans.* **31** 1095–105

[33] Sadava D E, Heller H C, Orians G H, Purves W K and Hillis D M 2006 *Life: The Science of Biology* (Sunderland, MA : Gordonsville, VA: W. H. Freeman)

[34] Silverthorn D U 2006 *Human Physiology: An Integrated Approach* (San Francisco: Benjamin Cummings)

[35] Brown, Lemay and Bursten 2006 *Chemistry The Central Science* (Upper Saddle River, NJ: pearson prentice hall)

[36] Livingston E H and Kohlstadt I 2005 Simplified Resting Metabolic Rate—Predicting Formulas for Normal-Sized and Obese Individuals *Obes. Res.* **13** 1255–62

[37] Simone G de, Devereux R B, Daniels S R, Mureddu G, Roman M J, Kimball T R, Greco R, Witt S and Contaldo F 1997 Stroke Volume and Cardiac Output in Normotensive Children and Adults *Circulation* **95** 1837–43

[38] Batterham A M, George K P and Mullineaux D R 1997 Allometric scaling of left ventricular mass by body dimensions in males and females *Med. Sci. Sports Exerc.* **29** 181–6

[39] Simone G and Galderisi M 2014 *Allometric Normalization of Cardiac Measures: Producing Better, but Imperfect, Accuracy* vol 27

[40] Dewey F E, Rosenthal D, Murphy D J, Froelicher V F and Ashley E A 2008 Does Size Matter? *Circulation* **117** 2279–87

[41] Chirinos J A, Segers P, Buyzere M L D, Kronmal R A, Raja M W, Bacquer D D, Claessens T, Gillebert T C, John-Sutton M S and Rietzschel E R 2010 Left Ventricular Mass *Hypertension* **56** 91–8

[42] Schalla S, Nagel E, Lehmkuhl H, Klein C, Bornstedt A, Schnackenburg B, Schneider U and Fleck E 2001 Comparison of magnetic resonance real-time imaging of left ventricular function with conventional magnetic resonance imaging and echocardiography *Am. J. Cardiol.* **87** 95–9

[43] Hubka M, Bolson E L, McDonald J A, Martin R W, Munt B and Sheehan F H 2002 Three-dimensional echocardiographic measurement of left and right ventricular mass and volume: in vitro validation *Int. J. Cardiovasc. Imaging* **18** 111–8

[44] Armstrong A C, Gidding S, Gjesdal O, Wu C, Bluemke D A and Lima J A C 2012 LV Mass Assessed by Echocardiography and CMR, Cardiovascular Outcomes, and Medical Practice *JACC Cardiovasc. Imaging* **5** 837–48

[45] Devereux R B and Reichek N 1977 Echocardiographic determination of left ventricular mass in man. Anatomic validation of the method. *Circulation* **55** 613–8

[46] Camarda J, Saudek D, Tweddell J, Mitchell M, Woods R, Otto M, Simpson P, Stendahl G, Berger S and Zangwill S 2013 MRI validated echocardiographic technique to measure total cardiac volume: A tool for donor–recipient size matching in pediatric heart transplantation *Pediatr. Transplant.* **17** 300–6

[47] Sorabella R A, Guglielmetti L, Kantor A, Castillero E, Takayama H, Schulze P C, Mancini D, Naka Y and George I 2015 Cardiac Donor Risk Factors Predictive of Short-Term Heart Transplant Recipient Mortality: An Analysis of the United Network for Organ Sharing Database *Transplant. Proc.* **47** 2944–51

[48] Johnson E J and Goldstein D 2003 Do Defaults Save Lives? *Science* **302** 1338–9

[49] Khan A M, Green R S, Lytrivi I D and Sahulee R 2016 Donor Predictors of Allograft Utilization for Pediatric Heart Transplantation *Transpl. Int.* n/a-n/a

[50] Anon UNOS | Working together. Saving lives.

[51] Renlund D G, Taylor D O, Kfoury A G and Shaddy R S 1999 New UNOS rules: historical background and implications for transplantation management *J. Heart Lung Transplant.* **18** 1065–70

[52] West L J, Pollock-Barziv S M, Dipchand A I, Lee K J, Cardella C J, Benson L N, Rebeyka I M and Coles J G 2001 ABO-Incompatible Heart Transplantation in Infants *N. Engl. J. Med.* **344** 793–800

[53] Taghavi S, Wilson L M, Brann S H, Gaughan J and Mangi A A 2012 Cardiac Transplantation Can Be Safely Performed With Low Donor-to-Recipient Body Weight Ratios *J. Card. Fail.* **18** 688–93

[54] Kanani M, Hoskote A, Carter C, Burch M, Tsang V and Kostolny M 2012 Increasing donor-recipient weight mismatch in pediatric orthotopic heart transplantation does not adversely affect outcome *Eur. J. Cardiothorac. Surg.* **41** 427–34

[55] Patel N D, Weiss E S, Nwakanma L U, Russell S D, Baumgartner W A, Shah A S and Conte J V 2008 Impact of Donor-to-Recipient Weight Ratio on Survival After Heart Transplantation Analysis of the United Network for Organ Sharing Database *Circulation* **118** S83–8

[56] Zafar F, Rossano J W, Price J F, Denfield S W, Heinle J S and Morales D L 2010 Listing Pediatric Patients on the Heart Transplant Waiting List with Weight Ranges Limits Donor Pool Unnecessarily *J. Surg. Res.* **158** 174

[57] Hosenpud J D, Pantely G A, Morton M J, Norman D J, Cobanoglu A M and Starr A 1989 Relation between recipient: donor body size match and hemodynamics three months after heart transplantation *J. Heart Transplant.* **8** 241–3

[58] Chan B B K, Fleischer K J, Bergin J D, Peyton V C, Flanagan T L, Kern J A, Tribble C G, Gibson R S and Kron I L 1991 Weight is not an accurate criterion for adult cardiac transplant size matching *Ann. Thorac. Surg.* **52** 1230–6

[59] Reichart B 1991 Size matching in heart transplantation. *J. Heart Lung Transplant. Off. Publ. Int. Soc. Heart Transplant.* **11** S199-202

[60] Constantine M, Harold H, Jon B. K, Laman A. G J, Brian L. G, Samual R. W, Francisco E and Larry N. C 1988 Infant orthotopic cardiac transplantation. *J. Thorac. Cardiovasc. Surg.* **96** 912–24

[61] Fullerton D A, Gundry S R, Alonso de Begona J, Kawauchi M, Razzouk A J and Bailey L L 1992 The effects of donor-recipient size disparity in infant and pediatric heart transplantation. *J. Thorac. Cardiovasc. Surg.* **104** 1314–9

[62] Gulshan K. S, Philip L, Luis J. R, Casey H, Micheal S. M, Samuel B and Jack G. C 1993 Clinical significance of weight difference between donor and recipient in heart transplantation. *J. Thorac. Cardiovasc. Surg.* **106** 444–8

[63] Morley D, Boigon M, Fesniak H, Brubaker P, Walter J, Fitzpatrick J, Chojnowski D, Smith A, Alpern J and Brozena S 1993 Posttransplantation hemodynamics and exercise function are not affected by

body-size matching of donor and recipient *The Journal of heart and lung transplantation* International Society for Heart and Lung Transplantation. Annual meeting and scientific sessions vol 12 (Elsevier) pp 770–8

[64] Fukushima N, Gundry S R, Razzouk A J and Bailey L L 1995 Growth of oversized grafts in neonatal heart transplantation *Ann. Thorac. Surg.* **60** 1659–64

[65] Verbraecken J, Van de Heyning P, De Backer W and Van Gaal L 2006 Body surface area in normal-weight, overweight, and obese adults. A comparison study *Metabolism* **55** 515–24

[66] Costanzo-Nordin M R, Liao Y, Grusk B B, O'Sullivan J E, Cooper R S, Johnson M R, Siebold K M, Sullivan H J, Heroux A H, Robinson J A and Pifarre R 1990 Oversizing of donor hearts: beneficial or detrimental? *J. Heart Lung Transplant. Off. Publ. Int. Soc. Heart Transplant.* **10** 717–30

[67] Razzouk A J, Johnston J K, Larsen R L, Chinnock R E, Fitts J A and Bailey L L 2005 Effect of oversizing cardiac allografts on survival in pediatric patients with congenital heart disease *J. Heart Lung Transplant.* **24** 195–9

[68] Blackbourne L H, Tribble C G, Langenburg S E, Sinclair K N, Rucker G B, Chan B B K, Spotnitz W D, Bergin J D and Kron I L 1994 Successful use of undersized donors for orthotopic heart transplantation—with a caveat *Ann. Thorac. Surg.* **57** 1472–6

[69] Tamisier D, Vouhé P, Le Bidois J, Mauriat P, Khoury W and Leca F 1996 Donor-recipient size matching in pediatric heart transplantation: a word of caution about small grafts. *J. Heart Lung Transplant. Off. Publ. Int. Soc. Heart Transplant.* **15** 190–5

[70] Tjang Y S, Stenlund H, Tenderich G, Hornik L, Bairaktaris A and Körfer R 2008 Risk Factor Analysis in Pediatric Heart Transplantation *J. Heart Lung Transplant.* **27** 408–15

[71] Bayoumi A S, Liu H and Fynn-Thompson F 2013 Donor-Recipient Size Matching in Pediatric Heart Transplantation: Is Weight the Most Appropriate Parameter To Predict Outcomes in All Age Groups? *J. Heart Lung Transplant.* **32** S128–9

[72] Dipchand A I, Edwards L B, Kucheryavaya A Y, Benden C, Dobbels F, Levvey B J, Lund L H, Meiser B, Yusen R D, Stehlik J and International Society of Heart and Lung Transplantation 2014 The registry of the International Society for Heart and Lung Transplantation: seventeenth official pediatric heart transplantation report--2014; focus theme: retransplantation *J. Heart Lung Transplant. Off. Publ. Int. Soc. Heart Transplant.* **33** 985–95

[73] Mather P J, Jeevanandam V, Eisen H J, Piña I L, Margulies K B, McClurken J, Furakawa S and Bove A A 1995 Functional and morphologic adaptation of undersized donor hearts after heart transplantation *J. Am. Coll. Cardiol.* **26** 737–42

[74] Conway J, Chin C, Kemna M, Burch M, Barnes A, Tresler M, Scheel J N, Naftel D C, Beddow K, Allain-Rooney T, Dipchand A I and The Pediatric Heart Transplant Study Investigators 2013 Donors' characteristics and impact on outcomes in pediatric heart transplant recipients *Pediatr. Transplant.* **17** 774–81

[75] Tang L, Du W, Delius R E, L'Ecuyer T J and Zilberman M V 2010 Low donor-to-recipient weight ratio does not negatively impact survival of pediatric heart transplant patients *Pediatr. Transplant.* **14** 741–5

[76] Ziaziaris W, Chew H C, Dhital K, Hayward C, Pleass H and Macdonald P 2014 Size and Gender Matching in Heart Transplantation – Optimizing Donor Utilization in an Era of Changing Donor and Recipient Characteristics *Curr. Transplant. Rep.* **1** 266–72

[77] de Simone G, Daniels S R, Devereux R B, Meyer R A, Roman M J, de Divitiis O and Alderman M H 1992 Left ventricular mass and body size in normotensive children and adults: Assessment of allometric relations and impact of overweight *J. Am. Coll. Cardiol.* **20** 1251–60

[78] de Simone G, Devereux R B, Daniels S R, Koren M J, Meyer R A and Laragh J H 1995 Effect of growth on variability of left ventricular mass: Assessment of allometric signals in adults and children and their capacity to predict cardiovascular risk *J. Am. Coll. Cardiol.* **25** 1056–62

[79] George K, Sharma S, Batterham A, Whyte G and Mckenna W 2001 Allometric analysis of the association between cardiac dimensions and body size variables in 464 junior athletes *Clin. Sci.* **100** 47–54

[80] Sluysmans T 2005 Theoretical and empirical derivation of cardiovascular allometric relationships in children *J. Appl. Physiol.* **99** 445–57

[81] Zuckerman W A, Richmond M E, Singh R K, Chen J M and Addonizio L J 2012 Use of height and a novel echocardiographic measurement to improve size-matching for pediatric heart transplantation *J. Heart Lung Transplant.* **31** 896–902

[82] Parry A and Large S 1994 Donor-recipient size match in heart transplantation *J. Thorac. Cardiovasc. Surg.* **108** 1150–1

[83] Hahn E, Zuckerman W A, Chen J M, Singh R K, Addonizio L J and Richmond M E 2014 An Echocardiographic Measurement of Superior Vena Cava to Inferior Vena Cava Distance in Patients <20 Years of Age With Idiopathic Dilated Cardiomyopathy *Am. J. Cardiol.* **113** 1405–8

[84] Patel A, Bock M J, Wollstein A, Nguyen K, Malerba S and Lytrivi I D 2016 Donor–recipient height ratio and outcomes in pediatric heart transplantation *Pediatr. Transplant.* **20** 652–7

[85] Canter C E 2016 Fitting Heart Transplantation to Adults With Congenital Heart DiseaseSquare Peg in a Round Hole?∗ *J. Am. Coll. Cardiol.* **68** 918–20

[86] Canter C E and Kantor P F 2007 Heart transplant for pediatric cardiomyopathy *Prog. Pediatr. Cardiol.* **23** 67–72

[87] Chen J M 2014 Heart Transplant: Transplantation for Congenital Heart Disease *Oper. Tech. Thorac. Cardiovasc. Surg.* **19** 30–46

[88] Montalvo J and Bailey L L 2010 Operative Methods Used for Heart Transplantation in Complex Univentricular Heart Disease and Variations of Atrial Situs *Oper. Tech. Thorac. Cardiovasc. Surg.* **15** 172–84

[89] Boucek M M, Aurora P, Edwards L B, Taylor D O, Trulock E P, Christie J, Dobbels F, Rahmel A O, Keck B M and Hertz M I 2007 Registry of the International Society for Heart and Lung Transplantation: Tenth Official Pediatric Heart Transplantation Report—2007 *J. Heart Lung Transplant.* **26** 796–807

[90] Jonas R A 2014 *Comprehensive Surgical Management of Congenital Heart Disease, Second Edition* (Boca Raton: CRC Press)

[91] Park S S, Sanders D B, Smith B P, Ryan J, Plasencia J, Osborn M B, Wellnitz C M, Southard R N, Pierce C N, Arabia F A, Lane J, Frakes D, Velez D A, Pophal S G and Nigro J J 2014 Total artificial heart in the pediatric patient with biventricular heart failure *Perfusion* **29** 82–8

[92] Moore R A, Madueme P C, Lorts A, Morales D L S and Taylor M D 2014 Virtual implantation evaluation of the total artificial heart and compatibility: Beyond standard fit criteria *J. Heart Lung Transplant.* **33** 1180–3

[93] Moore R A, Madueme P C, Lorts A, Morales D L and Taylor M D 2015 Virtual Implantation of the 50cc Total Artificial Heart *J. Heart Lung Transplant.* **34** S89

[94] Antretter H and Laufer G 2001 Surgical Techniques for Cardiac Transplantation *Acta Chir. Austriaca* **33** 17–24

[95] Kadner A, Chen R H and Adams D H 2000 Heterotopic heart transplantation: experimental development and clinical experience *Eur. J. Cardiothorac. Surg.* **17** 474–81

[96] Cooper D K C, Novitzky D, Becerra E and Reichart B 1986 Are there Indications for Heterotopic Heart Transplantation in 1986? *Thorac. Cardiovasc. Surg.* **34** 300–4

[97] Bellumkonda L and Bonde P 2012 Ventricular assist device therapy for heart failure--past, present, and future *Int. Anesthesiol. Clin.* **50** 123–45

[98] Hetzer R, Alexi-Meskishvili V, Weng Y, Hübler M, Potapov E, Drews T, Hennig E, Kaufmann F and Stiller B 2006 Mechanical Cardiac Support in the Young With the Berlin Heart EXCOR Pulsatile Ventricular Assist Device: 15 Years' Experience *Semin. Thorac. Cardiovasc. Surg. Pediatr. Card. Surg. Annu.* **9** 99–108

[99] Morgan J A and Edwards N M 2005 Orthotopic Cardiac Transplantation: Comparison of Outcome Using Biatrial, Bicaval, and Total Techniques *J. Card. Surg.* **20** 102–6

[100] Dreyfus G, Jebara V, Mihaileanu S and Carpentier A F 1991 Total orthotopic heart transplantation: an alternative to the standard technique *Ann. Thorac. Surg.* **52** 1181–4

[101] Aleksic I, Czer L S C, Freimark D, Takkenberg J J M, Dalichau H, Valenza M, Blanche C, Queral C A, Nessim S and Trento A 1996 Resting Hemodynamics After Total Versus Standard Orthotopic Heart Transplantation *Thorac. Cardiovasc. Surg.* **44** 193–8

[102] Bouchart F, Derumeaux G, Mouton-Schleifer D, Bessou J P, Redonnet M and Soyer R 1997 Conventional and total orthotopic cardiac transplantation: a comparative clinical and echocardiographical study. *Eur. J. Cardio-Thorac. Surg. Off. J. Eur. Assoc. Cardio-Thorac. Surg.* **12** 555–9

[103] Aziz T M, Burgess M I, El-Gamel A, Campbell C S, Rahman A N, Deiraniya A K and Yonan N A 1999 Orthotopic cardiac transplantation technique: a survey of current practice *Ann. Thorac. Surg.* **68** 1242–6

[104] Weiss E S, Nwakanma L U, Russell S B, Conte J V and Shah A S 2008 Outcomes in Bicaval Versus Biatrial Techniques in Heart Transplantation: An Analysis of the UNOS Database *J. Heart Lung Transplant.* **27** 178–83

[105] Austin P C and Steyerberg E W 2015 The number of subjects per variable required in linear regression analyses *J. Clin. Epidemiol.* **68** 627–36

[106] Redlarski G, Palkowski A and Krawczuk M 2016 Body surface area formulae: an alarming ambiguity *Sci. Rep.* **6** 27966

[107] Keys A, Fidanza F, Karvonen M J, Kimura N and Taylor H L 2014 Indices of relative weight and obesity *Int. J. Epidemiol.* **43** 655–65

[108] Montgomery D C, Peck E A and Vining G G 2015 *Introduction to Linear Regression Analysis* (John Wiley & Sons)

[109] Wilson W J 1998 *Regression Analysis: Statistical Modeling of a Response Variable* (San Diego: Academic Press)

[110] Williams K, Thomson D, Seto I, Contopoulos-Ioannidis D G, Ioannidis J P A, Curtis S, Constantin E, Batmanabane G, Hartling L and Klassen T 2012 Standard 6: Age Groups for Pediatric Trials *Pediatrics* **129** S153–60

[111] Kemna M, Albers E, Bradford M C, Law S, Permut L, McMullan D M and Law Y 2016 Impact of donor–recipient sex match on long-term survival after heart transplantation in children: An analysis of 5797 pediatric heart transplants *Pediatr. Transplant.* **20** 249–55

[112] Smits J M, Thul J, De Pauw M, Delmo Walter E, Strelniece A, Green D, de Vries E, Rahmel A, Bauer J, Laufer G, Hetzer R, Reichenspurner H and Meiser B 2014 Pediatric heart allocation and transplantation in Eurotransplant *Transpl. Int.* **27** 917–25

[113] Sullivan G M and Feinn R 2012 Using Effect Size—or Why the P Value Is Not Enough *J. Grad. Med. Educ.* **4** 279–82

[114] Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, Tarassenko L and Mant D 2011 Normal ranges of heart rate and respiratory rate in children from birth to 18 years: a systematic review of observational studies *Lancet* **377** 1011–8

[115] Finley J P and Nugent S T 1995 Heart rate variability in infants, children and young adults *J. Auton. Nerv. Syst.* **51** 103–8

[116] Bushberg J T, Seibert J A, Jr E M L and Boone J M 2011 *The Essential Physics of Medical Imaging, Third Edition* (Philadelphia: LWW)

[117] Fogel M A 2010 *Principles and Practice of Cardiac Magnetic Resonance in Congenital Heart Disease: Form, Function and Flow* (John Wiley & Sons)

[118] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning: with Applications in R* (Springer)

[119] R Core Team 2017 *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing)

[120] Behnke A R, Feen B G and Welham W C 1995 The Specific Gravity of Healthy Men: Body Weight + Volume as an Index of Obesity *Obes. Res.* **3** 295–300

[121] Rathbun E N, Pace N, Hinshaw E and Buntin H 1945 Studies on body composition. 1. The determination of total body fat by means of the body specific gravity. *J. Biol. Chem.* **158** 667–76

[122] Calcagno V and Mazancourt C de 2010 glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models *J. Stat. Softw.* **34**

[123] Calcagno V 2013 *glmulti: Model selection and multimodel inference made easy*

[124] Burnham K P and Anderson D R 2013 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (New York, NY: Springer)

[125] Burnham K P and Anderson D R 2004 Multimodel Inference: Understanding AIC and BIC in Model Selection *Sociol. Methods Res.* **33** 261–304

[126] Burnham K P and Anderson D R 2001 Kullback-Leibler information as a basis for strong inference in ecological studies *Wildl. Res.* **28** 111–9

[127] Burnham K, Anderson D and Huyvaert K 2011 AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons *Behav. Ecol. Sociobiol.* **65** 23–35

[128] Tan M Y J and Biswas R 2012 The reliability of the Akaike information criterion method in cosmological model selection *Mon. Not. R. Astron. Soc.* **419** 3292–303

[129] Zuur A, Ieno E N, Walker N, Saveliev A A and Smith G M 2009 *Mixed Effects Models and Extensions in Ecology with R* (Springer Science & Business Media)

[130] Breusch T and Pagan A R 1979 A simple test for heteroscedasticity and random coefficient variation

[131] Jose Pinheiro and Douglas Bates and Saikat DebRoy and Deepayan Sarkar and {R Core Team} 2017 *{nlme}: Linear and Nonlinear Mixed Effects Models*

[132] Faraway J J 2016 *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition* (CRC Press)

[133] Mayer D G and Butler D G 1993 Statistical validation *Ecol. Model.* **68** 21–32

[134] Burrill D F 2007 Modeling and Interpreting Interactions in Multiple Regression

[135] Armstrong J S and Collopy F 1992 Error measures for generalizing about forecasting methods: Empirical comparisons *Int. J. Forecast.* **8** 69–80

[136] Hyndman R and Koehler A 2006 Another look at measures of forecast accuracy *Int. J. Forecast.* **22** 679–88

[137] Goodwin P and Lawton R 1999 On the asymmetry of the symmetric MAPE *Int. J. Forecast.* **15** 405–8

[138] Fox J 2000 *Multiple and Generalized Nonparametric Regression* (Thousand Oaks, Calif: SAGE Publications, Inc)

[139] Kutner M, Nachtsheim C and Neter J 2004 *Applied Linear Regression Models- 4th Edition with Student CD* (Boston; Montreal: McGraw-Hill Education)

[140] Koenker R and Hallock K F 2001 Quantile Regression *J. Econ. Perspect.* **15** 143–56

[141] Cohen J 1988 *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, N.J: Routledge)

[142] Plasencia J D, Ryan J R, Park S S, Nigro J J, Frakes D H, Pophal S G and Zangwill S D 2017 The Virtual Heart Transplant - The Next Step in Size Matching for Pediatric Heart Transplantation *J. Heart Lung Transplant.* **36** S165

[143] Packard J M, Strutner L A, Melton R S and Ackerman I P 1958 Heart size in adolescents *Am. J. Cardiol.* **2** 170–8

[144] Maresh M M 1948 Growth of the Heart Related to Bodily Growth During Childhood and Adolescence *Pediatrics* **2** 382–404

[145] Mensah Y B, Mensah K, Asiamah S, Gbadamosi H, Idun E A, Brakohiapa W and Oddoye A 2015 Establishing the Cardiothoracic Ratio Using Chest Radiographs in an Indigenous Ghanaian Population: A Simple Tool for Cardiomegaly Screening *Ghana Med. J.* **49** 159–64

[146] Clarke S M, Griebsch J H and Simpson T W 2004 Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses *J. Mech. Des.* **127** 1077–87

[147] Smola A J and Schölkopf B 2004 A tutorial on support vector regression *Stat. Comput.* **14** 199–222

[148] Witten I, Frank E and Hall M 2011 *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition* (Burlington, MA: Morgan Kaufmann)

[149] Coe R 2002 It's the effect size, stupid: what effect size is and why it is important

[150] Olejnik S and Algina J 2000 Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations *Contemp. Educ. Psychol.* **25** 241–86

[151] Soper D 2018 Free Statistics Calculators: Home *Free Stat. Calc. Version 40*

APPENDIX A

LIST OF ACRONYMS

Clinical Acronyms:

BMI	**B**ody **M**ass **I**ndex
BSA	**B**ody **S**urface **A**rea
CT	**C**omputed **T**omography
DRBW	**D**onor-**R**ecipient **B**ody **W**eight
HHT	**H**eterotopic **H**eart **T**ransplant
Ht	**H**eigh**T** (i.e., body height)
HTx	**H**eart **T**ransplant
IRB	**I**nternal **R**eview **B**oard
MR	**M**agnetic **R**esonance
mTCV	**m**easured **T**otal **C**ardiac **V**olume
OHT	**O**rthotopic **H**eart **T**ransplant
PCH	**P**hoenix **C**hildren's **H**ospital
PHTS	the **P**ediatric **H**eart **T**ransplant **S**tudy (an organization)
pTCV	**p**redicted **T**otal **C**ardiac **V**olume
TCM	**T**otal **C**ardiac **M**ass
TCV	**T**otal **C**ardiac **V**olume
UNOS	the **U**nited **N**etwork of **O**rgan **S**haring (an organization)
Wt	**W**eigh**t** (i.e., body weight or body mass)*

** Wt, i.e., body weight, will be used interchangeably for body mass herein. The justification for terminology misuse is elaborated upon, in chapter 1.*

Statistical Acronyms:

AICc	**c**orrected **A**kaike **I**nformation **C**riterion
LOOCV	**L**eave-**O**ne-**O**ut **C**ross-**V**alidation
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercent Error
ME	**M**ean **E**rror
MPE	**M**ean **P**ercentage **E**rror
MSE	**M**ean **S**quare **E**rror
RMSE	**R**oot **M**ean **S**quare **E**rror
SE	**S**tandard **E**rror
St. Dev.	**St**andard **Dev**iation
VIF	**V**ariance **I**nflation **F**actor
1st-Qt	**First Q**uar**t**ile
2nd-Qt	**Second Q**uar**t**ile
3rd-Qt	**Third Q**uar**t**ile

Other Acronyms:

USA	**U**nited **S**tates of **A**merica

APPENDIX B

EFFECT SIZE AND SAMPLE SIZE NEEDED FOR TWO MEAN COMPARISON

When determining if two or more sample populations are statistically different the p-value is referenced, however, the effect size is an equally important metric. The work herein used Cohen's two-tail, non-directional *d* equation:

$$d = \frac{|Mean\ A - Mean\ B|}{\sigma} \qquad \text{(Equation B.1)}$$

to estimate the effect size between two populations in which σ is a standard deviation [141]. The standard deviation could be for population *A*, population *B*, or it could be the pooled standard deviation of both populations [141,149]. To estimate Cohen's *d* herein, a pooled standard deviation was used to account for (1) population size and (2) variance differences between the populations [149,150]:

$$\sigma_{pooled} = \sqrt{\frac{(N_A - 1){\sigma_A}^2 + (N_B - 1){\sigma_B}^2}{N_A + N_B - 2}} \qquad \text{(Equation B.2)}$$

The effect size scales the mean difference between populations relative to their variance and therefore quantifies the strength of the difference. Larger effect sizes indicate less overlap between the two populations' normal distributions – the reduced overlap supports the population difference is of practically importance [113,141]. Without the effect size the p-value loses practical meaning in what is statistically-significant [113]. In particular, the weakness of reporting only the p-value is any effect size > 0 can produce a p-value < 0.05, i.e., be found to be "statistically-significant", as long as the sample sizes are large enough [113]. It is important to recognize that it is the effect size and not the sample size that controls the overlap between the two populations – this highlights the danger of using the p-value as a standalone metric [113,141]. Providing both the effect size and p-value allows the reviewer to understand how strong a difference between two populations is and therefore determine if the difference is truly statistically-significant in the practical sense.

An additionally important use of effect size is approximating the sample size a study needs to determine if a difference between populations is statistically-significant in the practical sense. In determining the needed sample size, one would need to know (or guess) (1) the population variances and (2) either the effect size or population mean difference they would like to detect. Reworking the mathematical relationship between sample size and effect size could be used to analyze why a relationship was found or not found to be statistically-significant.

The author used this mathematical relationship between sample size and effect size to analyze the results in chapter 6. For the analysis, the author derived a relationship between sample size and mean differences in two key steps. First, the relationship between sample size and effect size for a two mean comparison was approximated using Dr. Danial Soper's online Student T-test sample size calculator [151]. The computed sample size results were then fitted to model the effect size needed to detect a given sample size, as shown in Figure B.1. Second, the fitted equation – with Cohen's *d* equation plugged in – was reworked to approximate the sample size to mean difference relationship. The final, reworked equation related the sample size to the populations' mean difference in a Student T-test comparison; the final equation used was as follows:

$$Mean\ Difference = \sigma * \left(\frac{Sample\ Size}{22.792}\right)^{\frac{-1}{1.952}} \qquad \text{(Equation B.3)}$$

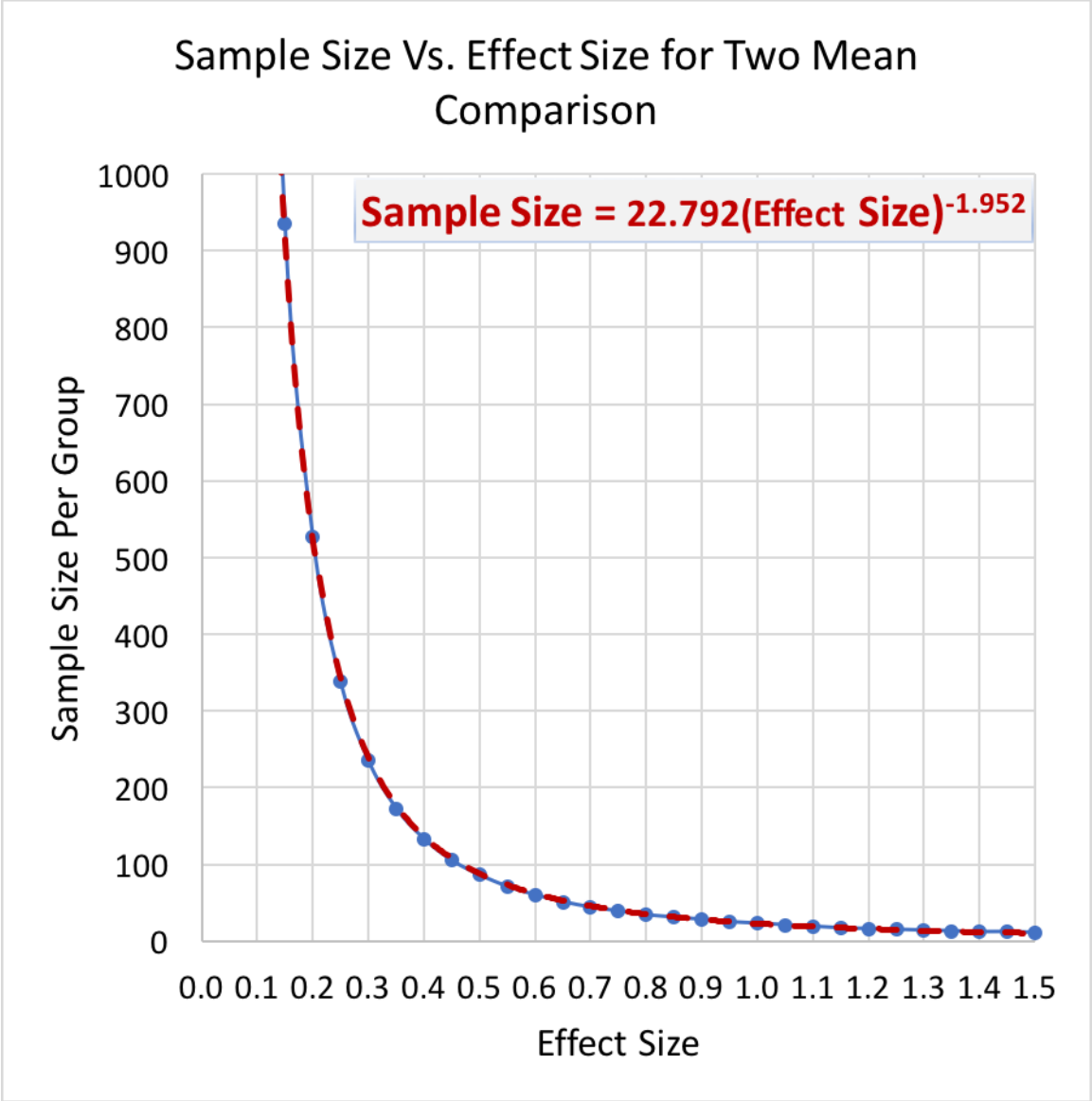

Figure B.1: The sample sizes needed (per sample population, i.e., group) for detecting a given effect size were approximated with a set alpha and power of 0.05 and 0.90, respectively. The relationship between the approximated sample size (blue; dots) vs. effect size fitted a power law function (red; dashed line).

APPENDIX C

DELEGATION OF TCV MODEL DEVELOPMENT AND VALIDATION

Jonathan Plasencia Model Contributions:

Plasencia developed Model A under the mentorship of Dr. Kamarianakis. Plasencia pointed out the potential clinical needed to develop a more conservative model after Model B was developed.

Yiannis Kamarianakis, Ph.D. Model Contributions:

Dr. Kamarianakis developed Model B and B* after reviewing Plasencia's preliminary Model A results. Model B* was developed in recognition that a more conservative model might be needed to prevent under-predictions. Furthermore, Dr. Kamarianakis helped mentor Plasencia as he developed Model A.

APPENDIX D

COPYRIGHT PERMISSIONS

Permissions for: Figure 6.3

Please see screen capture of permissions feedback obtained from automated system:

Permissions for: Figure 6.3

Figure D.1: Permission was not needed as (1) Plasencia is the author, (2) for Plasencia's thesis, and (3) not for commercial resale. Information was acquired by going to the "Get Rights and content" link on the ScienceDirect database while searching for the publication.