Secondary Usage of Electronic Health Record Data for Patient-Specific Modeling

Graham Karam Kim

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

John H. Gennari, Chair

Brian E. Carlson

Daniel L. Cook

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Secondary Usage of Electronic Health Record Data for Patient-Specific Modeling

Graham Karam Kim

Chair of the Supervisory Committee:
John H. Gennari
Biomedical Informatics and Medical Education

Translational research has become an important bridge that moves findings from basic science research to patients' bedside and to the clinical community. Unfortunately, this notion of translational research seems to be unidirectional in that basic research is translated into clinical research and practice, but basic science research does not seem to benefit as much from clinical medicine. In my dissertation, I leverage the availability of retrospective EHR data and use them with biosimulation models to translate data from clinical medicine to benefit biosimulation modeling. Biosimulation models are mathematical representations of biological systems, and they can help with mechanistic understanding of physiology and predict the dynamics of a biological system. Using clinical data with biosimulation models has the potential to benefit both the biosimulation modelers, as well as clinicians. The abundance of retrospective clinical data available for research is a promising alternative to the traditional method of validating models by

conducting resource-intensive prospective studies. These models can then be made patient-specific to simulate the physiology of individuals. When used in the clinical setting, these patient-specific models have the potential to be used by clinicians to better understand the underlying pathophysiology of the patient. In my research, I first conduct a scoping review of models in the literature to quantify model reproducibility and discover an appalling lack of model source code availability in publications. Then using a published hemodynamics model, I demonstrate using retrospective clinical dataset from right heart catheterizations to optimize and validate the model without needing to conduct burdensome prospective studies and explore potential clinical applications of patient-specific modeling. Finally, I describe an ontological approach to extend the data-model connection to be systematic and scalable. I demonstrate this approach by connecting cardiology data and lab results data with a hemodynamics model and several nephrology models, respectively.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincerest thanks to my mentor and advisor, John Gennari, for his guidance and undying support throughout the years. I would also like to thank Brian Carlson, Dan Cook, and Max Neal for their insight and encouragement. I would like to acknowledge the National Institute of Health for funding my graduate research. Finally, special thanks to my family and friends, especially my parents, my brothers, and my Biomedical and Health Informatics friends. Without their love and laughter, I would not be where I am today.

Chapter 1. Introduction

Translational research has become an important bridge that moves findings from basic science research to patient bedside and to the clinical community. Unfortunately, this notion of translational research seems to be unidirectional in that basic research is translated into clinical research and practice, but basic science research does not seem to benefit as much from clinical medicine.

Yet, there is an increasing volume of clinical data being captured via electronic health records (EHRs) and clinical data repositories that could benefit basic research. While the clinical data may not be collected for the purpose of being translated for basic science research, researchers could benefit from the tremendous volume of data from real human subjects.

My dissertation bridges this chasm between clinical medicine and basic science research by utilizing retrospective clinical data with computational physiology models to mitigate the burden on researchers to collect prospective physiological data. It leverages the abundance and ubiquity of electronic health record data, and the corpus of computational models available in model repositories and publications. As a result, biosimulation modeling benefits from being able to carry out model validation studies without conducting burdensome prospective data collection. Furthermore, clinical medicine can also benefit from the patient-specific models with the potential to aid clinical decision-making.

1.1 Motivation for Research

The motivation for this work is two-fold. First, there is the growing availability and prevalence of electronic health record (EHR) data. In fact, 84% of hospitals in the US have adopted at least a basic EHR system as of 2015 (Adler-Milstein et al., 2017). Even more astounding is the estimated 2,314 exabytes (1 exabyte = 1 billion gigabytes) of clinical data projected to be collected by 2020

(IDC, 2014). The potential for using retrospective EHR data to improve patient care and biomedical research has long been recognized, with potential applications including observational studies, surveillance, regulatory research, clinical research, and clinical analytics for quality improvement and cost reduction. Further accelerating the secondary usage of EHR are initiatives like Big Data to Knowledge (BD2K) (Ohno-Machado, 2014) and the creation of research-ready clinical data repositories that can accommodate clinicians, researchers, and administrators to self-query clinical data for different needs like cohort discovery, biomedical research, and quality improvement use, respectively.

At the same time, there is the notion of the *physiome*, with the goal of quantitatively describing the physiological dynamics and functional behavior of an organism. The Physiome Project of the International Union of Physiological Sciences (IUPS), for example, is an initiative to establish a computational modeling framework for the human body across multiple scales of granularity, incorporating biochemistry, biophysics, and anatomy of cells, tissues, and organs (Hunter et al., 2002). While the notion of physiome itself was conceived almost two decades ago, recent efforts in model curation, annotation, and model validation techniques have enabled an easier way to better utilize these mathematical models of physiology. The COmputational Modeling in BIology NEtwork (COMBINE) community, for example, has been combining various research groups and modeling standards under one roof for computational modeling practices that both meet individual needs, and promote interoperability (Neal et al., 2018a). Moreover, the recent inception of the Physiome Journal, with emphasis on reproducibility, reusability, and discoverability of mathematical and computational models, promotes this vision of *physiome* and further impowers computational physiology models to be utilized (Nickerson and Hunter, 2017).

The exact nomenclature for these mathematical representations of physiology varies. Different publications or communities refer to them as “computational models of biology,” “computational

physiology models," or "biosimulation models." Throughout my dissertation, I use these terms interchangeably to refer to the same thing: mathematical representations of biology that can simulate biological processes. In this dissertation, I will sometimes use "models" as a shorthand for computational physiology models.

One application for biosimulation models is patient-specific modeling. The goal of patient-specific modeling is to simulate the dynamics of tissues and organs of individual patients based on their patient-specific data. Patient-specific modeling has tremendous potential in biomedicine, not only as a research tool for understanding pathophysiology, but also as a clinical tool for improving clinical decision-making, predicting outcomes, and ultimately improving care for the patient. There is increasing interest and efforts to create patient-specific models, but one of the main challenges in patient-specific modeling is acquiring the data that is necessary to validate the model, to ensure that the modeler's hypothesis on the mathematical representation of pathophysiology does in fact hold true, or at least within an acceptable range, according to empirical measurements from a patient (Neal and Kerckhoffs, 2009). In fact, this data challenge exists not only for patient-specific modeling, but also for a wider range of models where the model must be validated against empirical data. Typically, these data are collected via prospective study where the modeler recruits human subjects and collects physiological data, or carries out wet lab benchwork to collect experimental data

To reduce the burden on the modelers of conducting prospective studies, I propose using the abundant clinical data that has already been collected from patients during the course of routine care. However, this raises the problem of matching relevant clinical data to corresponding model variables. One could certainly take an *ad hoc* approach to matching clinical data to models, but this approach is not scalable for the variety of clinical data, and the gamut of computational models available. For example, a modeler needing to validate his or her patient-specific model against

patient data would want to find the right dataset that corresponds to the model and its parameters. On the other hand, a clinical researcher with a clinical dataset wants to better understand the underlying physiology or conduct simulation studies. Currently for both cases, one needs to match clinical data and model parameters by sifting through model repositories, carefully reading the publication associated with the model, determining what physiological measures match with model parameters, then trying to determine which clinical datasets contain those parameters, if any. However, there is a vast number of computational physiology models available in both curated model repositories and generally in the literature. Furthermore, there is a wide gamut of poorly annotated clinical data with cryptic data field names that may not be readily available for perusing to determine best fit with the model of interest. This approach is clearly not scalable. There needs to be a more systematic method of matching clinical data with models.

1.2 Solution Approach and Scope

In order to bridge the gap between clinical data and computational physiology models, I propose a combination of optimization and ontological approach for using clinical data with computational models. First, I need to understand the landscape of computational physiology models, especially those in literature. While there are model repositories with manually curated models, they do not necessarily represent the body of models in publications that are publicly available. Thus, I carried out a scoping review of model publications available in literature, which I describe in section 1.2.1. Second, I need to be able to match data to model, but there are gaps in data where the model has more parameters than available data. Therefore, I need to employ optimization techniques to fill in these missing parameters, as described below in section 1.2.2. However, if I refer back to the *physiome* vision of reproducible, reusable, and discoverable models and simulations, the *ad hoc*

data-model matching is insufficient and unscalable. Finally, to make this approach more scalable, I need to leverage semantics, which I describe below in section 1.2.3.

1.2.1 Scoping Review of Models in Literature

In Chapter 3, I scope the status quo of computational physiology models in the literature to better understand the landscape of computational physiology models. I describe the search strategy, and the characterization of the resulting model publications by physiology being modeled, modeling paradigm, and their reproducibility. While there are curated model repositories, namely the Physiome Model Repository and BioModels Database, the availability of models directly from the literature is rather unclear. In this scoping review, I conducted three studies, each with increasing specificity with respect to *bona fide* computational physiology model publications.

In the first study, I searched for computational physiology models in PubMed using a combination MeSH terms. This primarily consisted of searching for publications that have been annotated with the MeSH terms, "Models, Biological," and "Computer Simulation." In addition to these MeSH terms, publication types were used to filter non-original model publications, such as review articles and meta-analyses. From this search, I sampled a subset of the search results and analyzed them to characterize the models by their biological domain, computational modeling paradigm, and modeling language or tool used. More importantly, I examined the publication for the availability of model code, model equations, and simulation parameters – either within the publication, supplemental material, or external hyperlink – for repeatability of the computational experiment and reproducibility of the biosimulation model.

Although the first study captured a wide range of models from different biological domains, none of the model publications analyzed from this search made the model source code available. In

order to increase the specificity of the search, I conducted a second study with a narrower search scope.

In the second study, I narrowed the scope of the search from all biosimulation models to just cardiovascular biosimulation models. While more specific than the first study with all biosimulation models, the cardiovascular biosimulation model search still returned no model publications that made model source code available.

In the third study, I took a different approach than querying PubMed. I examined a specific list of model publication from a review article on mathematical models of diabetes by Ajmera, et al. (Ajmera et al., 2013). While this list of publication may not be representative of all model publications in the literature, it did represent a very specific type of model, namely ordinary differential equation models of diabetes, with a very high proportion of publications fitting the inclusion criteria. In this list of model publications, there were finally two publications that made model source code available.

Very few model publications in this scoping review had model source code available, which is problematic for model reproducibility. Moreover, there is another issue: Curated models in repositories refer to the original publication, but those publications do not necessarily refer to the curated model code. In other words, even if a model publication might have a curated model code deposited in a centralized model repository, it may never be reachable from a literature search. Nonetheless, there are more recent publications that do include model source code in the supplemental materials section. Furthermore, there are recent publications that make great use of centralized model repositories, such as BioModels Database, in depositing their model in a standardized format, such as SBML, making their computational experiment repeatable and their model more accessible for reuse. Model availability and reproducibility are important concepts not only for this dissertation work, but also for the broader modeling community.

1.2.2 Parameterization and Optimization of a Model Using Patient Data

In Chapter 4, I describe optimizing a biosimulation model using a previously published model and retrospective clinical data. Currently, the biosimulation modeling process consists of a modeler building a mathematical model of some biological phenomenon, then validating the model against experimental data. For subcellular models, the validation step might consist of wet lab benchwork to collect the necessary data, and for clinical models this might require recruiting human subjects and carrying out experiments to collect physiological data necessary to validate the model. This process can be a major bottleneck for modelers.

My approach to optimizing and validating biosimulation models makes use of the already existing clinical data in lieu of conducting cumbersome prospective human subject experiments, thus drastically reducing the burden of the modelers on model validation. In Chapter 4, I demonstrate the feasibility of using patient data collected as a part of routine clinical workflow to parameterize an existing model for patient-specific modeling. With a better understanding of the model publication landscape from the work described in Chapter 3, I use a previously published model, and demonstrate model validation and patient-specific modeling using clinical data that has been collected during the normal course of clinical care.

As a proof-of-concept, I took a hemodynamics dataset from right heart catheterizations and parameterize a cardiovascular model that describes the hemodynamic properties and processes of blood, heart, and the vasculature. The clinical data, not being collected specifically for the purpose of validating this particular hemodynamics model, do not match with all of the model parameters. In fact, the model contains a greater number of parameters than the data elements available in the clinical dataset. In order to extrapolate from the limited data elements, I used model optimization techniques to estimate a carefully selected subset of parameters that are not explicitly measured in the clinical dataset. The result is a cardiovascular model that has been parameterized and optimized

for a set of hemodynamic measurements from a specific patient. Such patient-specific model could be used to estimate additional physiological values of the patient that is not directly measurable, such as the elastance of the vasculature. It could also be used to conduct simulation studies, or track the patient's trajectory over time.

1.2.3 Generalizing Data - Model Connection via Semantic Annotation

In Chapter 5, I describe a semantic approach to systematically, and precisely connect clinical data with model parameters. The data-model connection for model optimization described in Chapter 4 was a manual approach. While this manual approach for connecting one type of clinical dataset with one model is manageable, it is *not* scalable given the variety of clinical data available in different clinical data repositories, and the large corpus of biosimulation models available through curated model repositories, as well as models described in literature. Thus, the data-model parameter matching process should be systematic and scalable.

To generalize the matching of clinical data to biosimulation model parameters, I developed an approach for systematically connecting clinical data with computational physiology models via semantic annotation. This approach extends an existing biosimulation model annotation framework and tools to annotate clinical data in the same manner. To annotate clinical data, I used the *composite annotation* framework to precisely describe clinical measurements in the dataset.

The composite annotation framework uses multiple ontology terms to describe a biosimulation model variable. For example, there is no single ontology term that can fully describe a model variable that represents "right ventricular blood volume." Instead, the semantics of this variable can be decomposed into the physical property, "volume," and the physical entity "blood in the right ventricle."

There is already a body of biosimulation models whose variables have already been annotated with composite annotations such as the above. There is also a major push for model annotation in the biosimulation modeling community with initiatives like the Center for Reproducible Biological Models, whose goals include annotating biosimulation models to enhance reproducibility and reusability (Sauro et al., 2018). In my approach with clinical data, I used the same composite annotation framework to annotate clinical data and enable data interoperability and reusability with annotated biosimulation models.

To streamline the data annotation process, I leveraged SemGen, a model composition tool suite that includes an annotation module. The annotation module in SemGen is built around the notion of composite annotation. Since SemGen is not explicitly designed to annotate data, I converted the clinical data headers into a SemGen-readable format which can then be annotated using the existing tool. To help automate the annotation process, I developed a function in SemGen that can decompose a unit of measurement into its fundamental base units, which are then mapped to physical properties. Thus, SemGen can automatically annotate the physical property portion of the composite annotation given the unit of measurement for a data header.

1.3 Contributions

My research has two major contributions: 1) Supporting basic science in biosimulation modeling, and 2) augmenting clinical decision support with patient-specific modeling. For biologists and modelers, utilizing retrospective clinical data can accelerate the modeling process by providing an alternative to conducting burdensome and limited prospective experiments for model validation. For example, a biologist studying diabetes could create a computational model of insulin metabolism. Instead of having to recruit human subjects and performing phlebotomy to collect blood samples for a metabolic panel of chemicals in the subject's blood, the biologist could instead

source the data from a clinical data repository, where there might be decades worth of blood metabolic panel lab data from hundreds of thousands of patients and validate his or her model using these data. With my approach, translational research can benefit not just clinical research, but also basic science.

For clinicians, patient-specific models created by marrying biosimulation model with patient data has the potential to provide clinical insight. For example, a cardiology treating patients after a heart transplant could use a patient-specific hemodynamics model to simulate the patient's cardiac physiology. As further described in Chapter 4, such model could reveal physiological factors about the patient that would otherwise be impossible to measure directly, such as the resistance in the coronary arteries that might be indicative of cardiac allograph vasculopathy and ultimately transplant rejection. In another scenario, computational physiology models could use noninvasively collected physiological measurements to estimate physiological measurements that are traditionally collected via invasive methods. In my work, I used hemodynamics data from right heart catheter data to optimize a model, and one category of measurements that the model estimated was volumes of the heart chambers. However, if blood pressures and flows could be accurately estimated using volumetric measurements from noninvasive procedures like cardiac MRI or echocardiogram, the patient would not need to undergo invasive procedures like catheterization to collect the pressure and flow measurements.

1.4 Summary

The broad goal of my dissertation work is to leverage the abundance of existing clinical data and to use them with existing computation physiology models. My approach has three major components: 1) Reviewing the literature for model publications and assessing their model code availability, 2) using a published model and retrospective clinical data to demonstrate the

feasibility of optimizing and validating a model for patient-specific modeling, and 3) streamlining the clinical data and model matching process to make the use of retrospective clinical data for modeling more scalable and generalizable.

In my scoping review of model code availability from publications, only 2 out of 150 model publications examined made the model computational code available. This is a rather surprising and significant elucidation into the appalling state of model reproducibility in literature. Many modelers are keen on the importance of model reproducibility, and many are aware that model repeatability and reproducibility from publications is problematic. However, there is very little work quantifying the lack of model source code in literature. Furthermore, this scoping review described could better guide future efforts for the broader modeling community by describing the status quo of models in literature and highlighting some of the issues with model reproducibility, and the asymmetry of curated models referring to model publications but not vice versa.

Using such computational model available in literature, in conjunction with retrospective clinical data, I optimized and validated a computational physiology model with patient physiology data. This demonstrate the feasibility of optimizing and validating a model using retrospective data without needing to conduct burdening prospective studies that are currently used to collect data for model validation. In addition to the benefits for modelers, my approach also has important implications for clinicians. The patient-specific model can then be used to simulate the patient's physiology with the ability to estimate physiological measurements that are difficult to measure directly (e.g., elasticity of the aorta), track patient trajectories over time, and conduct perturbation studies. Furthermore, the retrospective data was collected during the course of normal clinical care from actual patients, demonstrating not only patient-specific modeling, but patient-specific modeling that could be more easily incorporated to clinical workflow without needing to collect additional patient data.

I extend the methodology described in Section 1.2.2 and Chapter 4 (optimization of biosimulation model using clinical data) to be generalizable by developing an informatics pipeline that leverages ontologies and semantic annotations. By annotating clinical data and matching them to model parameter annotations, the data-model matching process can be better automated, and with more semantic precision. More importantly, the broader potential for this work is building a knowledgebase of annotated clinical datasets in parallel to repositories of annotated biosimulation models. For modelers, finding suitable dataset to validate and simulate their models is a major hurdle. Finding suitable *clinical* dataset can be even more challenging due to the disparity of standards and controlled terminologies used in biosimulation modeling and clinical practice. My work bridges this gap by establishing an annotation process for clinical data that aligns with model annotations.

Chapter 2. Background

2.1 Biosimulation Models

A model is essentially a representation of something real. While a model can represent a variety of things in different formats, biosimulation models, or computational physiology models represent biological phenomena or biological systems. More specifically, these models represent physiological phenomena using mathematics.

So why do we bother with models? Modeling by definition takes a reductionist approach to reality — scoping reality into a tractable portion, reducing it into comprehensible parts, describing the relationship between those parts, and testing those relationships with observable outcome. As a result of translating reality into a model, information is lost. However, this reduction certainly has its benefits. In the context of biosimulation modeling, it helps us better understand biological phenomena by making explicit the various components and actors of biological processes and describing their relationships in the precise language of mathematics.

Computational physiology models have proven to be useful in a number of ways. These include better understanding of the mechanisms determining physiological function, conducting perturbation simulations, and predicting physiological trajectories over time. As these computational physiology models improve, they are better able to simulate patient-specific physiology with the ultimate goal of supporting clinical decision-making (Neal and Kerckhoffs, 2009).

These models are created based on theory and data, where theory guides the initial set of equations and parameters, and data are used in the crucial validation step to test the hypothesis set by the model. However, the model validation step is often costly and time-consuming as it requires prospective data collection. And yet, there is a wealth of underutilized retrospective clinical data.

One utility for these data is in computational physiology modeling, where retrospective clinical data has the potential to be a less costly source of data for model validation. Furthermore, to achieve the ultimate goal of supporting clinical decision-making in a patient-specific manner, these computational physiology models must be able to utilize the currently available clinical data and produce clinically relevant output.

There is a general consensus that computational physiology models can have a positive clinical impact with paradigms such as pharmacokinetic/pharmacodynamic (PK/PD) modeling or patient-specific modeling, but there is not any clear indication on the status quo of computational physiology models: How many of these models exist? What physiological phenomena do they model? Can models from the literature be reproduced by another modeler and be reused?

2.1.1 Modeling Standards

As biosimulation models have become more complex, solely relying on mathematics is insufficient to clearly describe the model. Furthermore, with the emphasis on reproducibility of these models and the computational experiments, we need standards to clearly define the format in which these models should be encoded in, as well as standards for how to annotate the model and the biology being described. In this section, I provide the background information on some of the existing modeling encoding standards and model annotation standards.

2.1.1.1 Physiome

The physiome is the quantitative and integrated description of an organism's physiological dynamics. Since its inception almost two decades ago, there have been numerous physiome projects with this common goal. Notably, there is the NSR Physiome Project based here in Seattle, which is an effort to define the physiome via the development of integrated quantitative and descriptive modeling (Bassingthwaighte, 2000). In addition, there is also the IUPS Physiome

Project for building a computational physiology modeling framework across scale, including biochemical, biophysical, and anatomical information on cells, tissues, and organs (Hunter and Borg, 2003). The IUPS Physiome Project also develops and maintains an XML-based modeling language, CellML, and a repository of CellML models, the Physiome Model Repository.

2.1.1.2 CellML and Physiome Model Repository

CellML is an XML-based modeling language, developed by the Auckland Bioengineering Institute, for describing biological models (Cuellar et al., 2003). Its purpose is to store and exchange computational models, and to facilitate better model reuse using model components. Model components are substructures within a model that encapsulates a portion of the model, allowing reuse of components from one model in another to accelerate model building.

CellML can be used to describe a wide range of biological phenomena, including sub-cellular biochemistry, to gross physiology, and it can describe the mathematics, typically algebraic equations or ordinary differential equations ODEs, and the model metadata, information about the model publication, authorship, and curation. The Physiome Project also maintains the Physiome Model Repository (PMR), which includes more than 500 curated CellML models (Yu et al., 2011) [available at: https://models.physiomeproject.org].

2.1.1.3 SBML and BioModels Database

Similar to CellML, SBML (Systems Biology Markup Language) is another XML-based modeling language (Hucka et al., 2003). It is commonly used to represent mathematical models of biochemical reactions, but it can also be used to encode models of cell signaling pathways, metabolic pathways, and gene regulation. Unlike CellML, SBML representation is typically limited to cellular and sub-cellular domain.

The SBML community also maintains a suite of software tools, including LibSBML API for working with SBML models, SBMLToolbox for working with SBML in MATLAB, MOCCASIN for translating ODE models in MATLAB to SBML. While the SBML Project does not itself produce models, the European Bioinformatics Institute (EMBL-EBI) maintains the BioModels Database (Le Novere et al., 2006; Li et al., 2010) [available at: http://www.ebi.ac.uk/biomodels-main/]. BioModels Database is a repository of computational models of biological processes that includes more than 600 curated models and more than 1,000 non-curated models encoded in SBML.

2.1.1.4 SemSim and SemGen

SemSim is a model-description architecture specifically designed to facilitate the sharing, reuse, and modular construction of biological models (Neal et al., 2009). The SemSim architecture is implemented in Web Ontology Language (OWL) (McGuinness et al., 2004), and the model is described with not only the mathematics, but also rich semantic knowledge.

SemGen is a tool that makes use of the SemSim model architecture to help automate modeling workflow of model visualization, annotation, extraction, and merge (Neal et al., 2015, 2018b). SemGen's annotator is particularly useful for my dissertation in annotating not only biosimulation models, but also clinical data. The annotator can convert CellML, SBML, and JSim (MML format) models into SemSim model format and annotate model variables using ontology terms. Moreover, the annotator can automatically determine the physical property and the corresponding Ontology of Physics for Biology term based on the unit of measurement of each model variable. In Chapter 5, I describe a method for converting clinical data into SemGen-compatible file format and annotating the dataset with SemSim-style composite annotations. Releases and source code for SemGen can be found on GitHub: https://github.com/SemBioProcess/SemGen/

2.1.1.5 Others

There are of course other modeling languages than the ones mentioned above. For example, there is JSim, a Java-based system for building and simulating numerical models and analyzing them (Butterworth et al., 2014). MATLAB (MATLAB Inc., Natick, MA. USA) is yet another commonly used modeling language in biosimulation. While MATLAB can be a powerful tool for numerical computing, its scope is not limited to biosimulation, and the MATLAB language does not have a strict format, giving modelers much freedom in how they describe a model. This can be useful for rapid model building, but it becomes quite problematic for model reuse. Since there is no standard format in which the model is written, there is no systematic method to parse the model code. Thus, reusing a model written in generic MATLAB code becomes a labor-intensive exercise.

In addition to these specialized modeling languages, modelers also use generic programming languages to represent their models. These include Python, C/C++, Fortran, and many more. All of these generic programming languages have a similar problem to generic MATLAB code in that model reuse can be very difficult due to the freedom the initial modeler has in writing the model, especially without proper documentation or annotation.

2.1.2 Model Annotation

In order for biosimulation models to be interoperable, reusable, and reproducible, they must be properly annotated with semantic precision using machine-readable knowledge resource terms. There are two levels of model annotation. One is the model-level annotation that describes the metadata about the model in general, including source publication, model authorship, and curatorial information. The other is model variable/parameter-level annotation that describes the physical and biological meaning of each model variable or parameter. In the scope of my dissertation, I focus on model variable/parameter-level annotation. While different model file

formats have different annotation syntax, most are limited to a single ontology term per model variable. However, a single ontology term is not sufficient to describe both the physics and the biology that a model variable or parameter represents. For example, there does not exist a single ontology term that can be used to annotate a model variable that represents blood volume in the right ventricle. Instead, one must use *composite annotations* to combine multiple ontologies to describe the physical property as well as the biological entity (Gennari et al., 2011). Continuing with the same example, blood volume in the right ventricle can be described with a composite annotation using both the Ontology of Physics for Biology (OPB) (Cook et al., 2008, 2011)and the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003, 2008):

OPB:Fluid volume <property_of> FMA:Portion of blood <part_of> FMA:Right ventricle

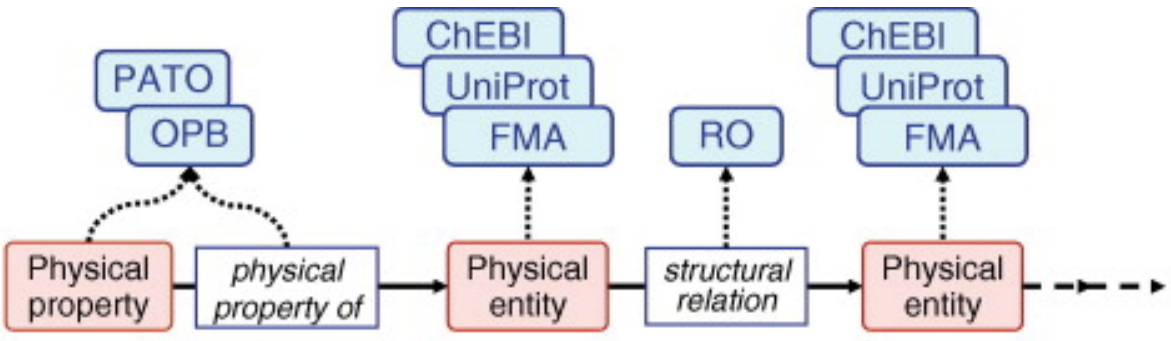

Figure 2.1 Structure of a composite annotation. Composite annotation uses multiple ontologies to describe both the physical property and the physical entities.

2.1.2.1 COMBINE Archive and OMEX Metadata

COmputational Modeling in BIology NEtwork (COMBINE) is an initiative that coordinates a variety of computational modeling communities and standards [https://co.mbine.org/]. In fact, the aforementioned CellML and SBML modeling formats are part of the COMBINE standards. One

of the standards defined by COMBINE is the COMBINE archive, which is a single file that bundles all of the files necessary to describe a computational model and its simulation experiment (Bergmann et al., 2014). This may include model files, data files, simulation experiment description, and annotation files, and this archive is encoded in the Open Modeling EXchange (OMEX) format.

Driven by the COMBINE community, there is a consensus to harmonize the semantic annotation in biosimulation modeling. The recent publication *Harmonizing semantic annotations for computational models in biology* describes the motivation and best practice recommendations for building a consensus approach to semantic annotation (Neal et al., 2018a). The recommendations are: 1) Encode annotations as RDF; use identifiers.org URI formatting and BioModels.net qualifiers. 2) Store annotations in a separate file. 3) Establish a dedicated group for developing a software library that supports semantic annotation standards. 4) Document which knowledge resources should be used for annotation and why. 5) Establish a repository of reusable annotations. 6) Ensure high-quality semantic annotations through training and quality control processes. 7) Establish and maintain collaborations with knowledge resource developers.

The OMEX Metadata specification is a document that formalizes the first two recommendations so that the community can adopt it as the standard for encoding and storing model annotations in the COMBINE archive. The specification is currently under work.

2.2 Electronic Health Record Data

Aside from biosimulation models, the other major component of my dissertation is electronic health record (EHR) data. 84% of hospitals in the United States had with at least a basic EHR system as of 2015 (Adler-Milstein et al., 2017). According to another research, the amount of electronic health record data in 2013 was 53 exabytes (1 exabyte = 1 billion gigabytes) and is

projected to be over 2,000 exabytes in 2020 (IDC, 2014). This increasing volume of EHR data presents a great opportunity for research use. Further elevating the opportunity of using EHR data for research use are initiatives like i2b2 (Informatics for Integrating Biology and the Bedside) (Murphy et al., 2010) that have led to the development of clinical data repositories, including research-ready de-identified clinical data repositories. Such repositories allow for much more efficient access to EHR data. The value of secondary usage of EHR data has been recognized for over a decade, with uses including quality assurance, public health surveillance, and clinical research (Hersh, 2007; Reis et al., 2017). In my dissertation, I demonstrate the secondary usage of EHR data with biosimulation models to explore a mechanistic, dynamical view of the relevant physiological systems and a patient-specific platform for clinical insight.

2.2.1 Usable EHR Data

While the raw amount of EHR data seems daunting, not all data are suitable for use with computational physiology models. The electronic health record captures a wide gamut of data, including both structured and unstructured data. Structured data includes data that have been captured in structured fields through a standardized data capture process. These include quantitative physiological measurements like vitals and lab results, but also non-physiological data like patient demographics and diagnoses. In the scope of my dissertation, I primarily focus on quantitative, structured, physiological data that are more conducive to use with computation physiology models.

Unstructured clinical data are typically qualitative data that are either captured in a non-standardized manner, or the information captured cannot be programmatically extracted. For examples, free-text narrative notes written by clinicians or unprocessed medical images fall under this category. By some industry estimates, 80% of EHR data is unstructured (Pak). With the

advancement of artificial intelligence (AI), natural language processing (NLP), and image segmentation, there are ongoing efforts to extract valuable information from unstructured clinical data and transform them into structured data that can be more readily used for research and analytics.

There is another category of underutilized clinical data. In my observation of the clinical workflow in the cardiac catheter lab, I noticed the catheter lab recording system was capturing a lot more data than what was actually recorded in the EHR. For instance, the catheter lab recording system records continuous measurements of the fluid pressure and its waveforms inside the patient's vasculature. Once the catheterization procedure is completed, the clinician interprets the waveforms and only records the summary hemodynamics data for the patient (e.g., mean arterial pressure, right ventricular end diastolic and end systolic pressure). Even though all of the continuous pressure waveforms were captured, only a few discrete data points are recorded in the medical record. For clinicians, perhaps only these summary data are of importance, but for researchers the raw data that is captured can be of tremendously rich source of information.

2.2.2 Clinical Standards

With regard to clinical data interoperability with computational models, there are two broad categories of clinical standards. First, there are *clinical terminologies* that represent the real-world meaning of clinical concepts. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (SNOMED International) is a comprehensive controlled vocabulary encompassing more than 300,000 unique concepts. The scope of SNOMED CT is the entirety of the electronic health record, including standardized terms ranging from anatomical structures to clinical measurement and diagnoses. SNOMED CT also defines a mechanism for creating

composites by combining multiple concepts into a SNOMED CT expression. For example, pneumococcal pneumonia in the lungs can be written as a SNOMED CT expression:

|Pneumococcal pneumonia| (SCTID: 233607000) |Finding site| (SCTID: 363698007) |Lung structure| (SCTID: 39607008)

Another commonly used clinical terminology standard is Logical Observation Identifiers Names and Codes (LOINC) (Forrey et al., 1996). LOINC is used for identifying laboratory tests, establishing a common language and codes for identifying health measurements and observations.

The other category of clinical standards important for data interoperability is *clinical information models* that define the structure and semantics of clinical information. openEHR and Fast Healthcare Interoperability Resource (FHIR) are examples of clinical information model standards. openEHR is an information model for EHR data that specifies how health data should be represented, processed, and visualized, rather than attempting to define the underlying concepts *per se* (Kalra et al., 2005). In Chapter 5, I describe how the openEHR platform can be used to map between standards and biosimulation modeling standards. While the use of FHIR is not explicitly explored in my dissertation, the API-based exchange of health information that FHIR enables is very promising for systematically getting patient data for computational modeling (FHIR Specification).

2.3 Summary

In this chapter, I have described the important concepts in biosimulation modeling and clinical data. In the following chapters, I build on these concepts and established knowledge, and describe my work in quantifying the model reproducibility from publications, connecting clinical data and biosimulation models, and creating a generalizable and scalable data-model connection methodology.

Chapter 3. Scoping Review of Computational Physiology Models for their Reproducibility

3.1 Background and Motivation

Public model repositories, such as the Physiome Model Repository (PMR) and BioModels Database, are vital resources for accessing computational models of biological processes. There is broad recognition that these resources provide value for scientists aiming to build from others' works. Both repositories are manually curated: Curators identify a publication with a model, and then work to develop and establish reproducibility of the code. These curators manually supplement models described in the literature with cross references and model source code, so the models can be reproduced and reused by other modelers. While these curated model repositories provide tremendous value for the modeling community, this method does not scale well with the pace of model publication.

Furthermore, it is unclear to what extent these repositories capture the models available in literature. Without these third-party curators, what percentage of models described in the literature are reproducible? More fundamentally, what percentage of publications about models include some access to, or information about the model code?

In order to elucidate the status of model availability and reproducibility in literature, I conducted a scoping review to characterize computational physiology models in literature. I looked at whether or not (a) the model code is available, (b) the modeling language used is stated, and (c) the equations and parameters used in the model are listed. I examined three categories of model publications, beginning from broad and going to narrow. First, using a combination of MeSH terms, I searched PubMed for computational physiology models broadly—this resulted in over 6,500 publications, of which only a fraction was relevant. Next, I searched more specifically for

cardiovascular models in PubMed—which returned over 1,000 publications. Finally, I examined 96 diabetes model publications identified in a diabetes modeling review article (Ajmera et al., 2013) as "Clinical Models". From each of the three categories, I randomly sampled and screened publications for inclusion. I analyzed the content of 50 full-text publications from each of categories for model code availability. In this chapter, I describe the specific method used in the scoping review, and the resulting revelation on the appalling state of model reproducibility or lack thereof.

3.2 Scoping Review Methods

3.2.1 Review Question and Objective

Scoping reviews are relatively emergent approach to reviewing research evidence (Davis et al., 2009). They are used to contextualize knowledge, identify the current state of understanding for a given topic, and identify gaps in the existing literature (Anderson et al., 2008; Arksey and O'Malley, 2005). While less formal than systematic reviews, scoping reviews can be useful for getting a broad survey of the literature in areas with much uncertainty. The objective of this scoping review is to better elucidate the availability and reproducibility of computational physiology models in literature. I characterize them by their modeling paradigm, modeling language used, availability of model equations, availability of simulation parameters, and availability of model source code used for simulation.

3.2.2 Search Strategy

I examined computational physiology models in three categories. First, I examine model publications over all biological and clinical domain searched and sampled from PubMed. Second, I focus on cardiovascular model publications searched and sampled from PubMed. And third, I

focus on diabetes model publications from a particular diabetes modeling review article, *The impact of mathematical modeling on the understanding of diabetes and related complications* by Ajmera, et al. The categories were chosen to gradually narrow the scope of models and increase the specificity of reviewing bona fide model publication.

The first category covers all biosimulation models available in PubMed. While this broad search provided a nice overview of models across all biological domains and modeling paradigm, it also returned a large number of publications describing models that did not fit the inclusion criteria, such as physical models (e.g., mannequins), signal processing models for MRIs and ECGs, and statistical models lacking mechanistic explanation of the biological phenomena.

In order to bolster the specificity of the search with model publications meeting the inclusion criteria, I examined a second category of model publications focusing solely on cardiovascular models by including the MeSH heading "Models, Cardiovascular," a subheading under "Models, Biological."

Finally, I examined diabetes model publications examined in a review article, *The impact of mathematical modeling on the understanding of diabetes and related complications* by Ajmera, et al. This analyzes mathematical modeling of glucose homeostasis, diabetic condition, and its associated complications. As such, the model publications discussed in this review article have already been identified as mathematical models, some of which have been coded into SBML models as indicated in the article.

3.2.2.1 Models from All Domains

Database searched: PubMed

Query used: "Models, Biological"[MH] AND "Computer Simulation"[MH] AND Humans[Mesh] NOT review[ptyp] NOT Meta-Analysis[ptyp] AND (Research Support, American Recovery and

Reinvestment Act[ptyp] OR Research Support, N I H, Extramural[ptyp] OR Research Support, U S Gov't, Non P H S[ptyp] OR Research Support, U S Gov't, P H S[ptyp] OR Research Support, U.S. Government[ptyp] OR Research Support, Non U S Gov't[ptyp] OR Research Support, N I H, Intramural[ptyp] OR Validation Studies[ptyp] OR Comparative Study[ptyp] OR Evaluation Studies[ptyp])

Date range: Up to May 11, 2018

Exclude: Statistical models, signal processing models, non-physiological models, non-computation models (e.g., simple mannequin), studies using previously published models without modification, studies comparing previously published models.

While the PubMed search query looks rather convoluted, the core of the query searches for publications annotated with MeSH terms for biological models and computer simulations for humans. The additional [ptyp] query parts screen for different publication types. Review articles and meta-analysis are explicitly excluded from the search since these articles analyze previously published models and does not describe an original model. The other [ptyp] query parts includes research articles that typically pertains to publication types that model publications are annotated with.

3.2.2.2 Models from Cardiovascular Domain

Database searched: PubMed

Query used: "Models, Cardiovascular"[MH] AND "Computer Simulation"[MH] AND Humans[Mesh] NOT review[ptyp] NOT Meta-Analysis[ptyp] AND (Research Support, American Recovery and Reinvestment Act[ptyp] OR Research Support, N I H, Extramural[ptyp] OR Research Support, U S Gov't, Non P H S[ptyp] OR Research Support, U S Gov't, P H S[ptyp] OR Research Support, U.S. Government[ptyp] OR Research Support, Non U S Gov't[ptyp] OR

Research Support, N I H, Intramural[ptyp] OR Validation Studies[ptyp] OR Comparative Study[ptyp] OR Evaluation Studies[ptyp])

Date range: Up to May 4, 2018

Exclude: Statistical models, signal processing models, non-physiological models, non-computation models (e.g., simple mannequin), studies using previously published models without modification, studies comparing previously published models.

Similar to the query used for the PubMed search for models across all biological domains, the query for cardiovascular model publications also use a combination of MeSH terms. The only difference between the all domain search and cardiovascular search is the first MeSH term, “Models, Cardiovascular” instead of “Models, Biological.” In fact, “Models, Cardiovascular” is a subheading under “Models, Biological,” thus the resulting publications returned from this query is a subset of the all domain search, focusing on models of the cardiovascular system.

3.2.2.3 Models from Diabetes Domain

For this category, I reviewed the 96 model publications listed as "Clinical Models" in the diabetes model review article by Ajmera, et al. Figure 3.1 lists these models as categorized by the authors of the review article. In this review article, the authors have categorized a corpus of mathematical models of diabetes as “clinical” and “non-clinical” based on the data used, the level of complexity, and the biological description. Within the “Clinical Models” category, they have further stratified the models by their purpose and physiological scale. In Figure 3.1, the four sub-categories of clinical diabetes models are: a) diagnosis; b) control; c) progression; and d) complications. The arrows in the figure indicates that a model was derived/adopted from the parent model. The lines indicate models representing similar biological phenomenon, but not derived from another model.

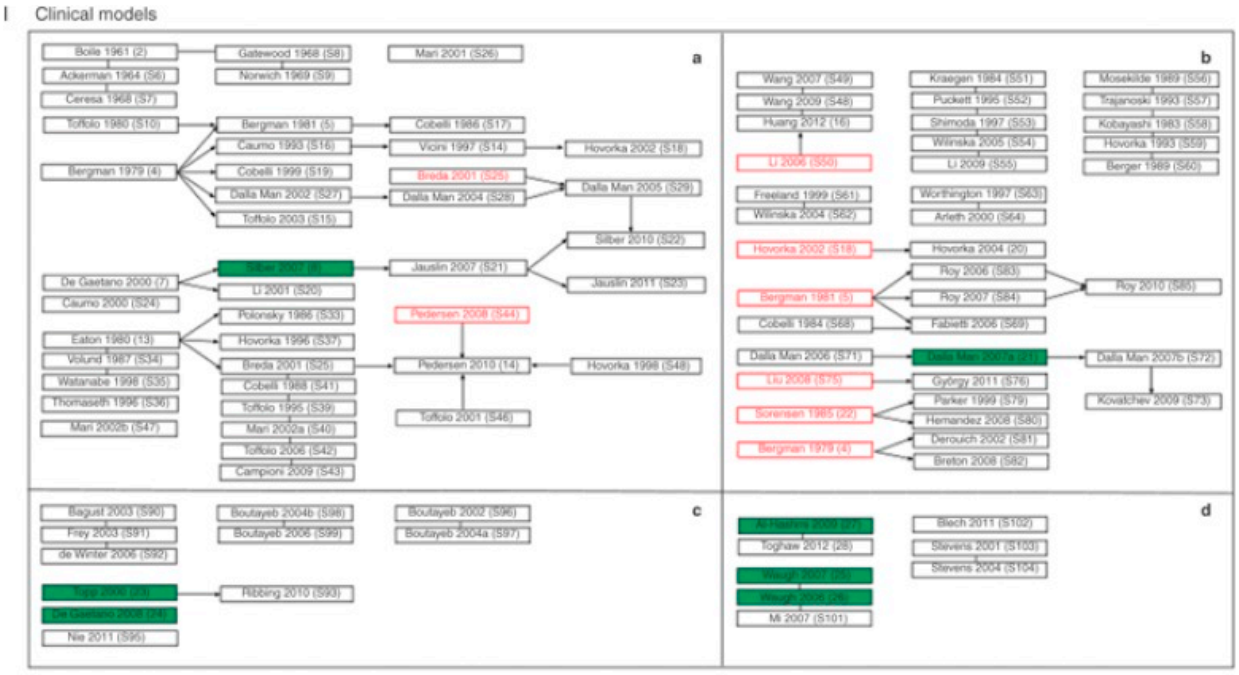

Figure 3.1. List of clinical diabetes model publications reviewed in Ajmera 2013.

3.3 Framework for Analysis

Biological domain: This only applies for the search results over all model domains, as the other two categories already have specified biological domains. This field indicates the broad category of biological phenomena that the model publication describes.

Modeling paradigm: This field describes the type of modeling paradigm used by the model, such as a system of ordinary or partial differential equations, or finite elements.

Modeling language: This field indicates the modeling tool or programming language used to code the model. Sometimes the publications describe both the software and language used to generate the model as well as code used for parameter estimation or curve fitting. Only the modeling language is indicated here.

Model equations available: Model publications were then analyzed for whether they listed a comprehensive list of model equations. Model equations are certainly more important for ODE and PDE models, but finite element and CFD models still require mathematical equations to describe the structural characteristics or flow dynamics.

Model parameters available: Similar to model equation availability, model publications were analyzed for whether they listed a comprehensive list of model simulation parameters. Oftentimes, publications resulting from model optimization, parameter estimation, or curve fitting. However, these parameters are not indicative of the initial conditions or parameters used to run the model simulation, thus resulting in very difficult, if not impossible, reproduction of the simulation described in the publication.

Model code available: This is perhaps the most important aspect of faithful and practical reproducibility of published models. For model code availability, publications were searched for the raw model/simulation code, as well as availability of the code in a remote repository via hyperlinks.

3.4 Results

From each of these categories, publications were randomly sampled until I identified 50 model publications that met the inclusion criteria. I conducted a detailed review on these 50 model publications from each category (150 total) and analyzed their model characteristics and reproducibility. The result of the analysis is listed below.

3.4.1 Results from All-Domain

Table 3.1 Scoping review results for the broad search for models in all biological domains.

Search	**Number of publications**	
Total publications retrieved from search	6909	
Number of publications randomly sampled	150	
Excluded from analysis	100	
Included for analysis	50	
Biological domain (Some model publications describe multiple domains)	**Number of publications**	**Percentage of total publications included for analysis**
Biomechanics	19	38%
CV	18	36%
Metabolism	6	12%
PKPD	4	8%
Electrophysiology	4	8%
Respiratory	3	6%
Diabetes	2	4%
Oncology	2	4%
Radiation	2	4%
Neural	1	2%
Model paradigm (ODE, PDE, etc.)		
Finite element	14	28%
ODE	12	24%
CFD	6	12%
PDE	5	10%
Population PK	3	6%
Others	10	20%
Model language (MATLAB, C++, etc.)	**Number of publications**	**Percentage of total publications included for analysis**
MATLAB	8	16%
ANSYS	5	10%
ABAQUS	4	8%
NONMEM	3	6%
C++	2	4%
Python	1	2%
Fortran	1	2%

SAAM II	1	2%
Others	10	20%
Not specified	18	36%
Reproducibility	**Number of publications**	**Percentage of total publications included for analysis**
Model equations listed	9	18%
Model equations NOT listed	41	82%
Simulation parameters listed	20	40%
Simulation parameters NOT listed	30	60%
Model code available	0	0%
Model code not available	50	100%

3.4.2 Results from Cardiovascular Domain

Table 3.2 Scoping review results for the search for cardiovascular models.

Search	**Number of publications**	
Total publications retrieved from search	1111	
Number of publications randomly sampled	96	
Excluded from analysis	46	
Included for analysis	50	
Model paradigm (ODE, PDE, etc.)	**Number of publications**	**Percentage of total publications included for analysis**
CFD	15	30%
Finite element	12	24%
ODE	8	16%
PDE	5	10%
Others	10	20%
Model language (MATLAB, C++, etc.)	**Number of publications**	**Percentage of total publications included for analysis**
ANSYS	13	26%
MATLAB	7	14%
C/C++	4	8%
ADINA	2	4%

COMSOL	1	2%
Maple	1	2%
SolidWorks	1	2%
LabVIEW	1	2%
Others	7	14%
Not specified	13	26%
Reproducibility	**Number of publications**	**Percentage of total publications included for analysis**
Equations listed	15	30%
Equations NOT listed	35	70%
Simulation parameters listed	15	30%
Simulation parameters NOT listed	35	70%
Model code available	0	0%
Model code not available	50	100%

3.4.3 Results from Diabetes Domain

Table 3.3 Scoping review results for the diabetes model publications from Ajmera, et al.

Search	**Number of publications**	
Total publications retrieved	96	
Number of publications randomly sampled	53	
Excluded from analysis	3	
Included for analysis	50	
Model paradigm (ODE, PDE, etc.)	**Number of publications**	**Percentage of total publications included for analysis**
ODE	34	68%
PDE	8	16%
Mixed effect	6	12%
Agent-based	1	2%
Matrix model	1	2%
Model language (MATLAB, C++, etc.)	**Number of publications**	**Percentage of total publications included for analysis**
MATLAB	8	16%
NONMEM	6	12%

XPP	2	4%
SAAM II	3	6%
C++	2	4%
Fortran	1	2%
BASIC	1	2%
Others	3	6%
Not specified	24	48%
Reproducibility	**Number of publications**	**Percentage of total publications included for analysis**
Model equations listed	36	72%
Model equations NOT listed	14	28%
Simulation parameters listed	28	56%
Simulation parameters NOT listed	22	44%
Model code available	2	4%
Model code not available	48	96%

3.4.4 Results Summary

Surprisingly, all but two model publication examined had no model code available. One was a journal publication while the other was described in a doctoral dissertation. A few publications referred to an external link with model source code that is no longer available). Furthermore, most model publications in the general and cardiovascular categories only listed a subset of equations and parameters used for model simulation. More than a third of model publications in the diabetes category only list a subset of the model equations and parameters. In this third category, about 7% of the models were included in BioModels library, but all of these were added and curated after publication. Thus, even for publications with curated models, a scientist simply reviewing the literature would have no easy way of finding these model codes.

3.5 Conclusion

Despite the push towards reproducibility of computational models, the vast majority of model publications do not provide sufficient information to reproduce the model simulations they

describe. At a minimum, modelers and authors should indicate where the source code is available along with some information about the language used so that their computational experiment can be reproduced by others. Ideally, this would include describing all of the simulation parameters used to produce the published results, and also submitting the model source code into centralized repositories such as BioModels Database or the PMR.

In the past, the printed publication medium may not have been conducive to including long lines of source code. However, with digital publication, online source code repositories like GitHub, and centralized model repositories, the sharing of source code is no longer an infrastructure issue. Perhaps there has been a publication cultural bias on the unimportance of raw code, but rather a heavier emphasis on results. Moving forward, one approach to ameliorate this situation would be if journals and publishers require or at least encourage authors to submit model source code and all of the equations and simulation parameter as part of the supplemental material for the publication.

Fortunately, I do think more and more emphasis is being placed on publishing the source code and data. Although they were not sampled and analyzed in my scoping review, there are recent submissions to the BioModels Database whose publications do in fact include model source in the supplemental material of the publication or by making explicit that the model code was deposited to BioModels Database. Furthermore, the Center for Reproducible Biomedical Modeling is partnering with journal publishers to not only encourage model code sharing but also help annotate models being published to make them more reusable (Sauro et al., 2018).

3.5.1 Limitations

While the paucity of model code from publication in this scoping review was quite appalling, this work does have its limitations. First, the literature searches described in section 3.2.2.1 and 3.2.2.2

are limited to PubMed searches. While PubMed searches across a wide range of biomedical journals, there might be other journals, especially those with a focus on mathematics or computer science such as Society for Industrial and Applied Mathematics journals, that are not index in PubMed. These journals might include publications on biosimulation modeling, and given the journal's focus on computation, the source code availability might differ from my results for PubMed.

Another limitation is that I was the only reviewer during this process. While the most important question of whether the model publication contained source code can be assessed with objectivity, other characteristics like the inclusion of equations and parameters had more room for interpretation. In fact, most model publications include at least some equations, but they did not always include a full set of equations necessary to reproduce the model. This process was even more challenging when equations were embedded in the text along with supporting equations that are not part of the model, *per se*, but used to derive another equation. Having a second reviewer independently assess these characteristics and then together reaching a consensus would have been a stronger analysis approach.

3.5.2 Future Work

There are several avenues of future work that could be explored stemming from this work. An obvious short-term work to bolster the current work would be to increase the sample size and perhaps include a second reviewer to confirm the characterization of model publications. Given a larger sample, an additional useful analysis would be to stratify the model publications by year to uncover any trends in model reproducibility. As mentioned before, I have anecdotally identified several recent publications making the model source code available and depositing the model to a model repository. Quantitative analysis of model code availability, especially with respect to the

start of efforts like the Center for Reproducible Biomedical Modeling, would serve as valuable evidence of the positive impact these efforts are making.

Chapter 4. Model Optimization Using Clinical Data

4.1 Introduction

In this chapter, I demonstrate the feasibility of using retrospective clinical data with computational physiology models. To do so, I use right heart catheterization (RHC) hemodynamics data, and parameterize the *Minimal haemodynamic system model including ventricular interaction and valve dynamics* (Smith et al., 2004). Using established computational techniques, I optimize the model to fit patient data from RHC, and I discuss the clinical and modeling implications of patient-specific modeling with EHR data.

4.1.1 Patient-Specific Modeling

While biosimulation models are often used as research tools, one of the more exciting and clinically applicable uses of biosimulation models is in patient-specific modeling. As the name suggests, patient-specific modeling simulates the individual physiology of a patient, and it has tremendous potential in biomedicine.

The notion of patient-specific modeling is not novel in itself. In fact, there have been numerous models that simulate patient physiology using patient-specific data. These studies prospectively recruit subjects and conduct procedures for data collection. While these studies are able to collect controlled data specific to the model in question, the procedures are costly and time-consuming.

For example, a study by Caroli, et al. validated a patient-specific computational vascular network model by conducting a multicenter, prospective clinical study to collect longitudinal data on arm vasculature before and after surgery (Caroli et al., 2013). In another case, a patient-specific model was simply validated against previously accepted literature data as described in *Patient-Specific Modeling of Blood Flow and Pressure in Human Coronary Arteries (Kim et al., 2010)*:

“The computed coronary flow and pressure and the aortic flow and pressure waveforms were realistic as compared to literature data.”

Unlike these previous models and model validation studies, I leverage the abundant clinical data stored in the EHR and use *existing* patient data to address some of the current challenges in burdensome data collection process. The main goal of this work is to validate the feasibility of using a limited set of retrospective patient data to create a patient-specific instantiation of a hemodynamic system model. One potential clinical implication of this work is estimating patients' physiological characteristics that are typically unmeasurable in a clinical setting, such as the elastance of the aorta or the pulmonary vasculature resistance. Another potential clinical application is in tracking these clinically unmeasurable patient-specific physiology longitudinally over time.

4.2 The Model

4.2.1 Model Selection

In order for a model to be made patient-specific, it has to check off a list of requirements. Firstly, it should be in a clinically relevant scale and domain. Many models, especially those in the BioModels Database, model sub-cellular biological phenomena, such as ion transport and electrophysiology. While these models are useful for better understanding and simulating subcellular physiology, they have less immediate clinical relevance, i.e., routine clinical workflow does not capture the data that such a model simulates, and the clinician would not be able to directly utilize these models to improve care for a patient.

Secondly, it needs to be simple enough. A complex model with a plethora of variables and equations, or a very detailed 3D model may require too much computational time. This is

especially problematic if the model were to be used at the point of care. In addition, for a model with numerous constitutive variables, i.e., variables with no physiological meaning such as curve-fitting variables, only a small portion of the model variables may be useful clinically. Also importantly, a model with a large number of parameters may not be uniquely identified with the relatively sparse clinical data, yielding an optimization open to too many degrees of freedom and resulting in potential overfitting and a physiologically less meaningful interpretation of the data.

Lastly, the model should yield clinical insight that cannot be derived otherwise. An example of a model that meets the first two requirements, but not the third is the *Creatinine kinetics and the definition of acute kidney injury* (Waikar and Bonventre, 2009). This model simulates kidney function using serum creatinine concentration to better define acute kidney injury. While this model is at a scale and domain of high clinical relevance and simplicity, it does not provide much insight about a patient's renal physiology. This model simulates the serum creatinine concentration over time and proposes new definitions for acute kidney injury. While serum creatinine level is an important proxy for renal function, there already exists a blood test that can directly measure serum creatinine level.

Given these constraints, I chose the cardiovascular model described in the publication, *Minimal haemodynamic system model including ventricular interaction and valve dynamics* by Smith, et al. (henceforth referred to as the "Smith model"). The Smith model was developed with the possibility of using clinical data for patient-specific modeling, and as such, it is a simple yet robust representation of hemodynamics in the cardiac chambers and the vasculatures.

4.2.2 The Smith Model

As previously stated, the Smith model was developed with the possibility of patient-specific modeling and providing rapid diagnostic feedback. Thus, the authors of the Smith model had the

following criteria for a minimal hemodynamics model: 1) Model parameters can be relatively easily determined or approximated for a specific patient; 2) The model can be run on a desktop computer in reasonable time; 3) Accurate prediction of trends; And 4) the full closed-loop model must be stable with minimal complexity and physiologically realistic inertia and valve effects.

The Smith model includes the hemodynamics in the heart, as well as the pulmonary and systemic circulations. The heart in the Smith model includes the two ventricles and their ventricular-ventricular interaction, but it does not include the atria in order to reduce the overall complexity of the model. The Smith model contains 41 input parameters including resistances and inertances of blood through the vasculature and valves, elastances of the vasculatures and ventricles, as well as end-diastole pressures and end-systole zero pressure volumes of the ventricles and the septum, and various constitutive parameters, such as exponent factors that determine the relationship between the ventricular volume and pressure during diastolic filling. Given these parameters, the model simulates output values for various blood flows, pressures, and volumes. Figure 4.1 illustrates a simplified schematics of the Smith model using analogous electrical circuit notations.

The Smith model has been encoded in CellML, and is available in the Physiome Model Repository at:

https://models.cellml.org/exposure/9d046663ba5cac5c8a61ac146183614b/smith_chase_nokes_shaw_wake_2004.cellml/view.

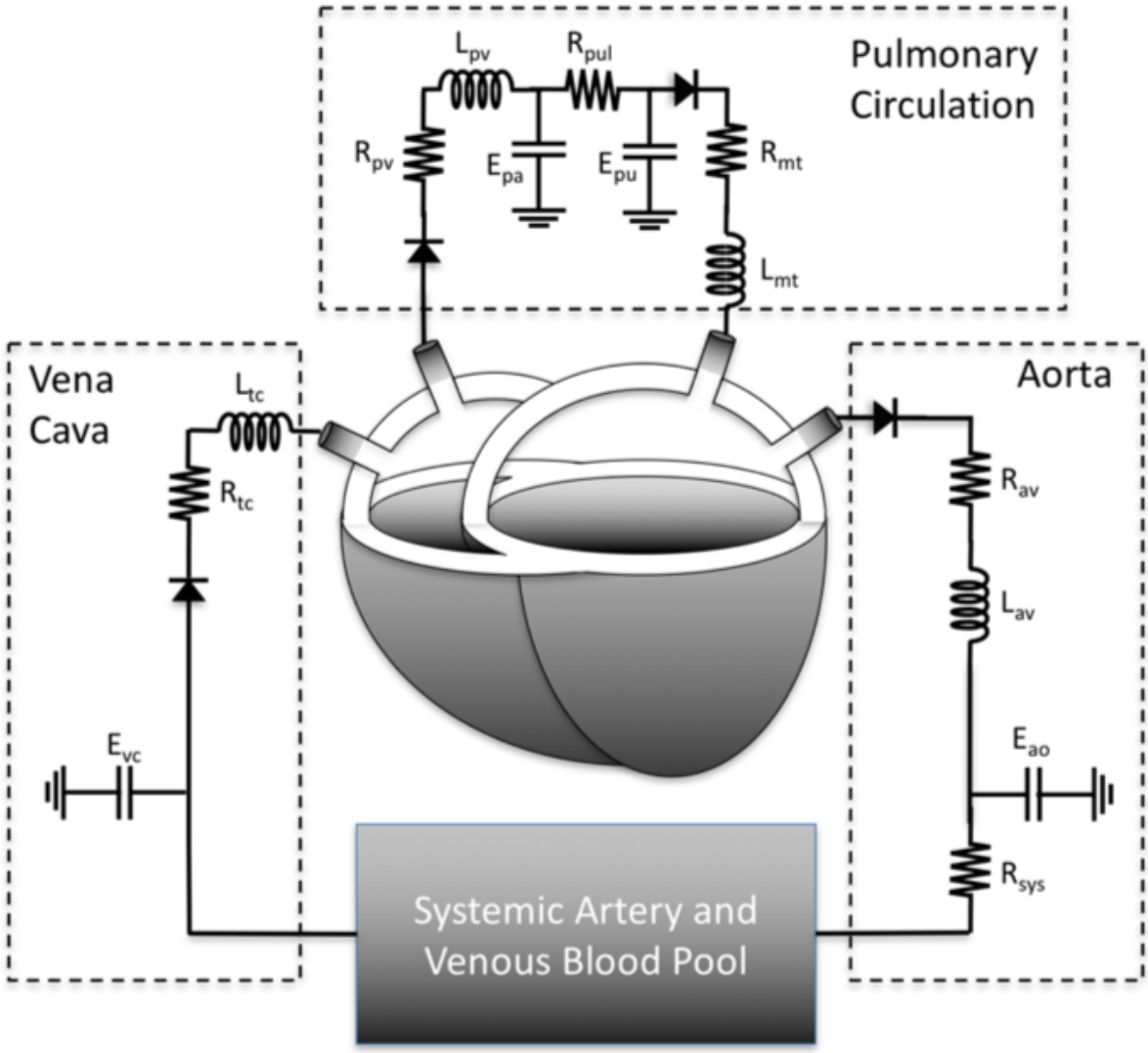

Figure 4.1 Schematic diagram of the Minimal haemodynamic system model. Smith, et al.

4.3 The Data

Given the model choice of the Smith hemodynamics model, I sought a suitable dataset that could be used to parameterize and validate the model. After consulting with several cardiologists, Drs. T. Dardas, P. Leary, and C. Masri, I decided to use the hemodynamics data that is collected during a typical right heart catheterization as a suitable data source for parameterizing the Smith model.

The right heart catheterization is an invasive procedure for obtaining hemodynamics data, during which a Swan-Ganz catheter is inserted into the heart via the jugular or another major vein and hemodynamics measurements are taken at the right atrium, right ventricle, and pulmonary artery. It is typically used to diagnose and manage patients with heart failure (HF), heart failure with preserved ejection fraction (HFpEF), and pulmonary hypertension (PH) (Patil et al., 2017). The RHC procedure is also routinely administered to heart transplant patients not only for hemodynamics measurements, but also for a heart tissue biopsy to monitor for rejection of the transplanted heart.

For this study, the data was limited to heart transplant patients with the rationale that a newly transplanted heart should be relatively healthy without any anatomical or physiological defects, such as ventricular septal defects or regurgitation that the Smith model does not accommodate for. One might raise the concern that a transplanted heart lacks baroreflex due to denervation. While this could be problematic in some models, the Smith model does not model baroreceptors, thus making data from transplant patients well suited.

I acquired the RHC data from the clinical data repository at the University of Washington Medicine Regional Heart Center. The RHC data in this repository was imported from the Mac-Lab Hemodynamic Recording System (GE Healthcare, Chicago, IL, USA) used in the UW cardiac catheterization lab. The repository was queried for RHC datasets from heart transplant patients with catheterization procedures ranging from March 6, 2014 to March 21, 2016. With an IRB approval, I received the dataset from the database administrator at the Regional Heart Center.

The datasets contain a minimum of the following twelve clinically measured values: right ventricular, pulmonary artery, and aortic pressures at diastole and systole; average pulmonary capillary wedge pressure, heart rate, cardiac output, body weight, height, and gender. The RHC dataset also includes calculated values of systemic vascular and pulmonary vascular resistance.

4.4 Model Parameterization and Optimization

In order to create a patient-specific version of the Smith model, the Smith model was parameterized and optimized using the data from RHC procedures. The optimization work was carried out in collaboration with Brian Carlson, a physiological modeling expert. All model code was developed, and optimization and simulation were performed in MATLAB (MATLAB Inc., Natick, MA. USA).

Comparing the parameters and variables simulated by the Smith model, and the data measured and recorded in the RHC procedure, I have identified 9 data points as suitable matches: RV systolic pressure, RV diastolic pressure, pulmonary arterial systolic pressure, pulmonary arterial diastolic pressure, pulmonary capillary pressure, aortic systolic pressure, aortic diastolic pressure, cardiac output, and average heart rate. In addition, I used height, weight, and gender from the patient record to estimate the total blood volume using Nadler's Formula (Nadler et al., 1962).

Most of these data points, with the exception of average heart rate, and total blood volume are model output measures. As such, I applied optimization techniques to adjust the model input parameters to produce model output measures that "fit" the clinical data. Given that the Smith model has 41 input parameters, adjusting all of these input parameters would both risk overfitting, as well as being too computationally burdensome.

4.4.1 Adjustable Parameter Set – Physiological Features

After consulting with cardiac modeling experts and cardiologist, I chose a minimal set of parameters from the Smith model to be adjustable in the optimization process to fit the RHC patient data. This parameter set was chosen based on their potential physiological and clinical significance, i.e., the adjustable parameters are not constitutive variables solely with mathematical purpose (e.g., curve-fitting parameters), but rather they represent properties of anatomical entities. Table 4.1 lists the adjustable parameter set chosen based on physiological features and their potential relevance

to cardiac pathophysiology. These parameters include various elastances and resistances, as well as the left ventricular free wall unstressed volume. The remaining parameters in the model were estimated from the clinical data where possible or set to the normal values used in the Smith model.

Table 4.1 A set of adjustable parameters selected based on physiological features to be optimized to fit the patient data.

Adjustable Parameters - Physiological	**Description**
$E_{es,lvf}$	LV Free Wall Elastance (mmHg/mL)
$V_{d,lv}$	LV Free Wall Unstressed Volume (mL)
$E_{es,rv}$	RV Free Wall Elastance (mmHg/mL)
$E_{es,pa}$	Pulmonary Artery Elastance (mmHg/mL)
$E_{es,pu}$	Pulmonary Vein Elastance (mmHg/mL)
R_{pul}	Pulmonary Vascular Resistance (mmHg*s/mL)
$E_{es,ao}$	Aorta Elastance (mmHg/mL)
R_{sys}	Systemic Vascular Resistance (mmHg*s/mL)
R_{pv}	Pulmonary Valve Resistance (mmHg*s/mL)
$E_{es,spt}$	Septum Wall Elastance (mmHg/mL)

While these adjustable parameters for optimization have physiological significance, they may not necessarily be easily identified in the Smith model. For example, the value for a parameter with

low sensitivity could be greatly altered without affecting the model output very much. In this case, the optimization result for an insensitive parameter could potentially vary widely, since the model output would not be affected as much as other more sensitive parameters. This may cause the optimization results to have physiologically nonsensical values, or inconsistent optimized values over different iterations of optimization. Furthermore, even after performing a sensitivity analysis, optimizing a parameter set with correlated parameters could also cause issues if the correlated parameters can negate each other's effect on the model output. In the next section, I report on the sensitivity analysis and correlation analysis, carried out with colleagues, as a more systematic approach to determine which parameters have the most impact on the output of the model, and which parameters are correlated by mathematical dependencies.

4.4.2 Adjustable Parameter Set – Sensitivity and Correlation Analysis

In collaboration with P. Woodall, et al., we conducted a sensitivity analysis and correlation analysis on a simplified Smith model to systematically identify a parameter set suitable for optimization with the RHC dataset (Woodall et al., 2018). In this work, we used a simplified version of the Smith model that does not include the ventricular-ventricular interaction, nor the inertances of blood through the valves.

Sensitivity analysis calculates how much each input parameter affects the model output in response to a perturbation. Furthermore, correlation analysis identifies which parameters are mathematically codependent using a sensitivity-based covariance analysis. The detailed mathematics and methods used for these analyses are described in Woodall, et al, 2018.

Using a sensitivity threshold of 0.01 and correlation threshold of 0.9, we identified the parameter set listed in Table 4.2. The sensitivity analysis identified diastolic filling exponents (numerical factors that describes the exponential relationship between the ventricular volumes and pressures

during diastole) in the left (λ_{lv}) and right ventricle (λ_{rv}) as most sensitive parameters in the reduced model. The sensitivity analysis identified left ventricular unstressed volume ($V_{d,lv}$) as not sensitive. In addition, correlation analysis found left ventricular elastance ($E_{es,lv}$) and right ventricular elastance ($E_{es,rv}$) to be correlated with the more sensitive left (λ_{lv}) and right ventricle (λ_{rv}) diastolic filling exponents, suggesting one set or the other can be identified but not both. Thus, the less sensitive ventricular elastances were not included in the parameter set in Table 4.2. While pulmonary valve resistance (R_{pv}) was identified as having low sensitivity, we included this parameter in the adjustable parameter set to provide maximum flexibility in optimizing right ventricular pressure, pulmonary arterial pressure, and cardiac output, which are available in the RHC dataset.

Table 4.2 A set of adjustable parameters selected based on sensitivity and correlation analysis to be optimized to fit the patient data.

Adjustable Parameters – Sensitivity Analysis	**Description**
λ_{lv}	LV diastolic filling exponent
λ_{rv}	RV diastolic filling exponent
$E_{es,pa}$	Pulmonary Artery Elastance (mmHg/mL)
$E_{es,pu}$	Pulmonary Vein Elastance (mmHg/mL)
R_{pul}	Pulmonary Vascular Resistance (mmHg*s/mL)
$E_{es,ao}$	Aorta Elastance (mmHg/mL)
R_{sys}	Systemic Vascular Resistance (mmHg*s/mL)
R_{pv}	Pulmonary Valve Resistance (mmHg*s/mL)

The optimization results and longitudinal analysis using the parameter set in Table 4.2 is further described in the manuscript under preparation with Woodall, et al. For the purpose of my dissertation, I describe the optimization results using the parameter set in Table 4.1, chosen based on physiological significance.

4.4.3 Optimization Techniques

I initially considered a gradient descent algorithm that iteratively steps towards the negative of the gradient to find the local minimum (Ruder, 2016). However, this required manually selecting initial values for each optimization. Thus, this method was not scalable for batch processing a large number of optimizations. Furthermore, the optimization results could vastly differ depending on the choice of initial values as the optimization algorithm was finding a local minimum near the initial values, rather than find the global minimum solution set. To address these concerns, I ultimately chose genetic algorithm optimization method, which stochastically samples the parameter space given only initial bounds.

A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution (Sivanandam and Deepa, 2008). The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution.

I performed the optimization on a shared scalable compute cluster for research (Hyak) at the University of Washington. The MATLAB optimization script was run on 1 node with 16 cores and 40 GB of memory. The runtime for the optimization widely varied depending on the initial seed population values for the GA, ranging from a few hours up to 20+ hours, sometimes timing

out on Hyak. Despite the varied optimization time, the GA was a better optimization algorithm than the gradient descent algorithm for finding a globally optimal solution set.

4.5 Optimization Results

RHC datasets from four patients were used to optimize the Smith model. Table 4.3 lists the optimized results of the 10 parameters chosen based on physiological features (Table 4.1) and their residuals, which indicates the error in the optimization result. The residual from these optimizations range between 0.79% to 4.2%, indicating that the optimized parameters result in output that fit relatively well to the patient RHC data. One interesting aspect of the optimization results is the variability of septum wall elastance between the patients. The optimization results would indicate patients 266 and 572 have very stiff septal walls, while patients 233 and 558 have more compliant septal walls. The extraordinarily stiff septal walls may indicate some pathophysiology causing the stiffening. However, it could also be a potential weakness of the optimization, especially if the septal wall elastance parameter has low sensitivity or correlation with another parameter.

Figures 4.2 through 4.5 show the simulation output for four patients. In subplot A, the right ventricular, aortic, pulmonary arterial, and pulmonary venous pressures are all plotted. Subplots D and E separates the pressures by left heart and right heart, respectively. Subplot B shows the left ventricular and right ventricular volumes. Subplots C and F combine the visualization of pressures and volumes into pressure-volume (PV) loops, with subplot C plotting the left ventricular PV loop, and subplot F showing the right ventricular PV loop.

Although clinician assessment would be needed to verify, these very different PV loops strongly suggest that these patients are in very different clinical states. For example, the PV loops for patients 266 and 572 show much more limited stroke volume in the right ventricle with elevated

end systolic volume compared to patients 233 and 558. This may be indicative of serious conditions like right heart failure. PV loops like these provide clinicians with snapshots of the patient's cardiac function, which can help them elucidate the underlying pathophysiology.

Table 4.3 Optimization results for each patient dataset.

Adjustable Parameter	**Patient 233**	**Patient 266**	**Patient 558**	**Patient 572**
LV Free Wall Elastance (mmHg/mL)	3.840	6.993	3.820	3.322
LV Free Wall Unstressed Volume (mL)	16.441	3.057	5.862	15.066
RV Free Wall Elastance (mmHg/mL)	0.858	0.197	0.510	0.328
Pulmonary Artery Elastance (mmHg/mL)	0.240	0.370	0.274	0.424
Pulmonary Vein Elastance (mmHg/mL)	0.221	0.624	0.238	0.717
Pulmonary Vascular Resistance (mmHg*s/mL)	0.0732	0.0957	0.0875	0.107
Aorta Elastance (mmHg/mL)	0.981	0.896	1.038	1.333
Systemic Vascular Resistance (mmHg*s/mL)	0.911	1.126	0.588	0.909
Pulmonary Valve Resistance (mmHg*s/mL)	0.0118	0.0162	0.00422	0.0142
Septum Wall Elastance (mmHg/mL)	11.972	101.44	2.858	106.157
Residual	0.00791	0.0244	0.04172	0.01438

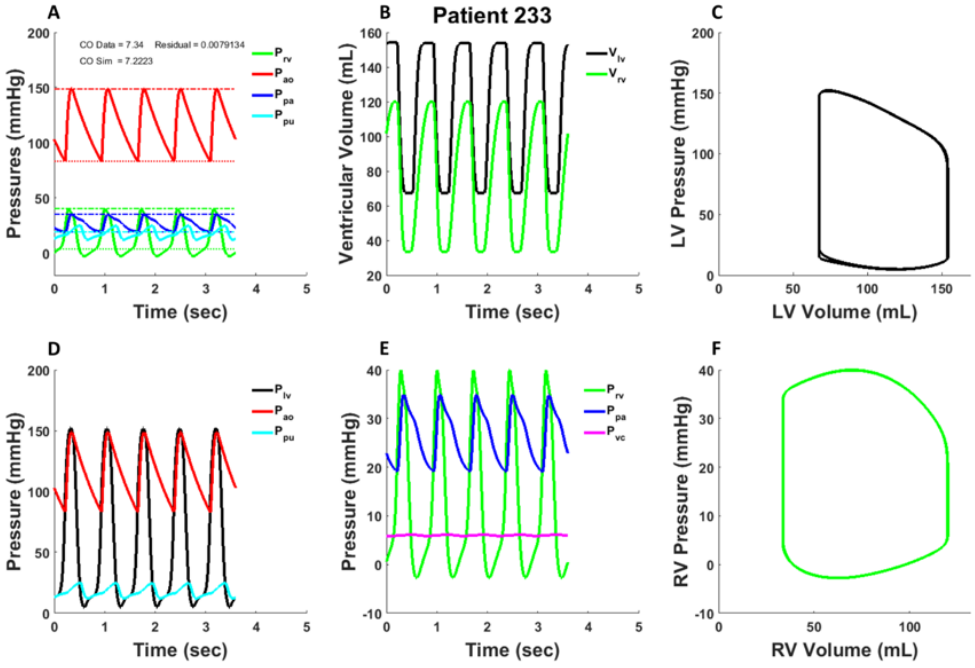

Figure 4.2 Simulation output of optimized model for patient 233.

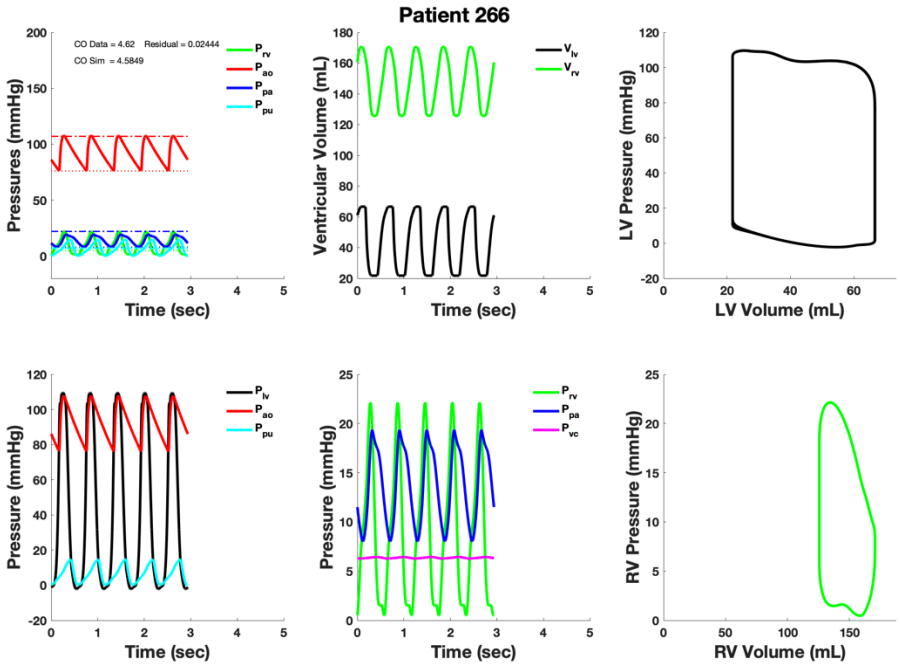

Figure 4.3 Simulation output of optimized model for patient 266.

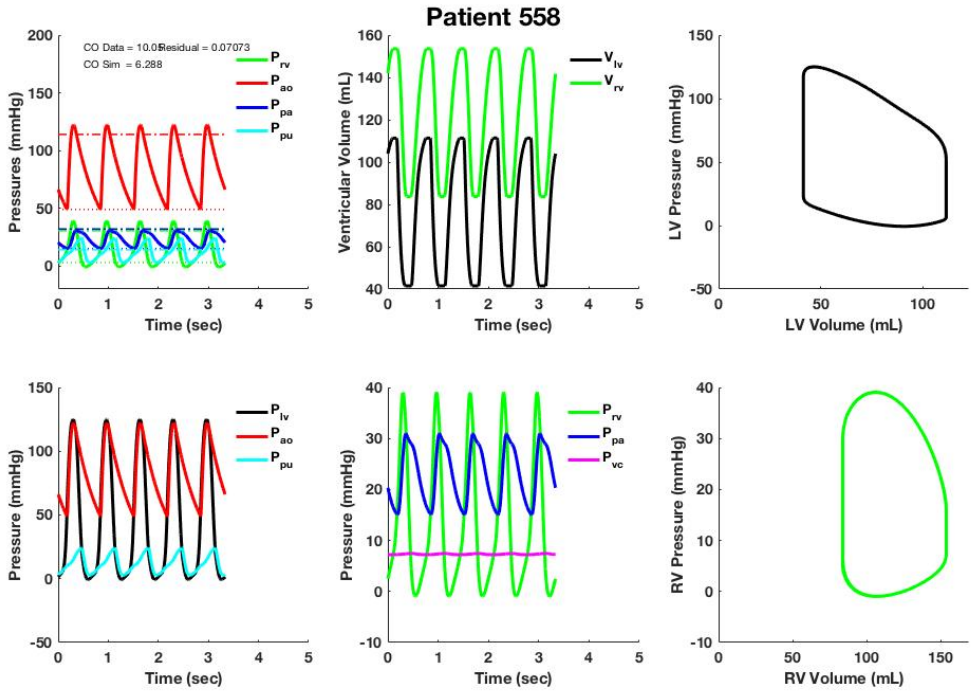

Figure 4.4 Simulation output of optimized model for patient 558.

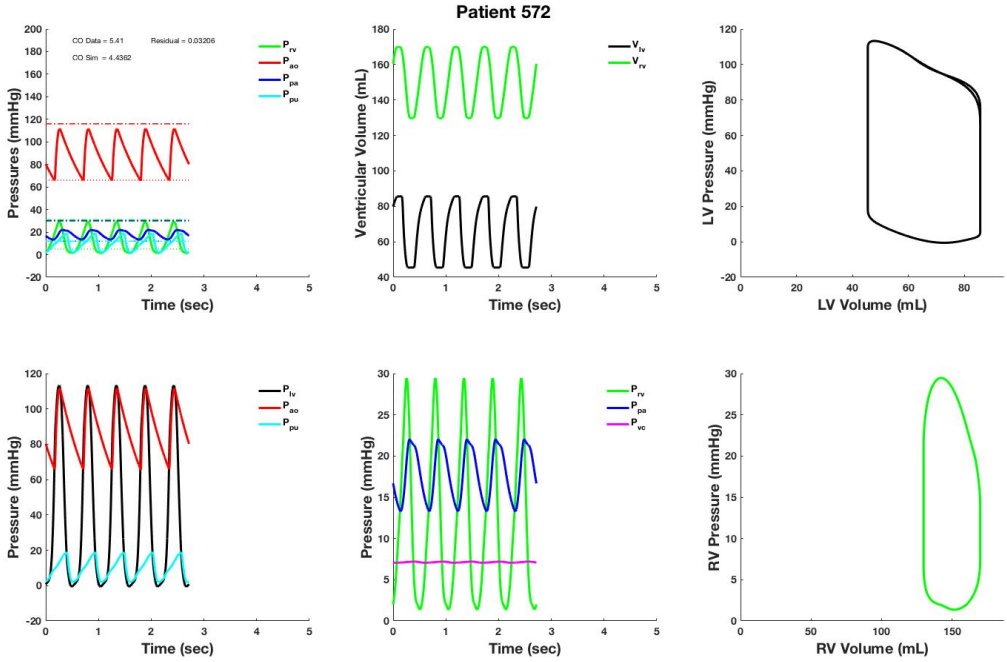

Figure 4.5 Simulation output of optimized model for patient 572.

4.6 Discussion and Significance

4.6.1 Clinical and Modeling Implications

There are several clinical benefits and implications associated with this work. Through optimization techniques, the model is able to estimate parameters about the patient that are clinically difficult, or even impossible to measure directly, such as the specific elastances of the aorta, pulmonary vein, pulmonary artery, left ventricular free wall, right ventricular free wall, and septal wall.

There are, however, accepted methods for estimating arterial stiffness (elastance), ranging from simple calculations based on pressures and volumes (Amin et al., 2011) to pulse wave velocity (PWV) measurements, which is the noninvasive gold standard for comparative measure of arterial stiffness (Townsend, 2017). Nonetheless, the model provides a more detailed and holistic characterization of the patient's physiology. The RHC procedure can only directly measure hemodynamic properties related to pressure and flow. Furthermore, the model simulates patient physiology as a function of time, whereas the clinically recorded measurements are generally limited to discrete data points or average values. Since the model is able to estimate volume parameters in addition to the pressure parameters over time, it can generate patient-specific pressure-volume loops, which can be used by clinicians to assess cardiovascular function, and even help diagnose difficult to diagnose conditions such as heart failure with preserved ejection fraction (HFpEF) in stable patients (Penicka et al., 2010).

Using multiple datasets over time for the same patient, the model could even be used to track the patient's physiology longitudinally, with the potential to highlight changes or abnormalities in the patient physiology that could lead to better prognosis and treatment. In fact, in collaboration with Woodall, et al., we have conducted sensitivity analysis on a cardiovascular system model based

on the Smith model to identify the more impactful parameters and optimize those parameters with longitudinal patient datasets to predict their cardiovascular function over time (Woodall et al., 2018). In this work, we optimized the modified Smith model using RHC data collected at several time points after the heart transplant for several patients. In this analysis, we were able to identify distinct trends in cardiac function for different patients, which could be used to suggest which patients might be having a less successful post-transplant recovery and warrant a closer monitoring.

In additional to clinical implications, this work also has modeling implications. Namely, this work demonstrates the feasibility of model parameterization and validating using existing clinical data. For modelers, this could mean real human data can be used to validate models without needing to conduct time-consuming and costly prospective studies, especially when recruiting for specific cohorts (e.g., heart transplant patients) that may be limited in sample size, or for data that may be difficult to collect (e.g., invasive right heart catheterization).

4.6.2 Limitations and Future Work

One limitation of this work is that RHC procedures are invasive. Using data collected via noninvasive procedures to estimate parameters that could only be directly measured by invasive procedures would establish far greater clinical benefits of using computational physiology models. In fact, one possible future work is to apply similar model optimization method using volumetric and cardiac output data collected via noninvasive modalities, such as echocardiogram and cardiac MRI, with the goal of estimating parameters, such as pressures that are currently collected via invasive procedures like right heart catheterization. This future work is further explored in the Future Directions section of Chapter 6.

Another limitation of this work is the lack of clinical validation. While the residuals from the optimized parameters indicate good mathematical fit, the parameter values and the simulation

results have not been thoroughly validated as a useful clinical tool. One potential clinical validation would be to compare the actual patient outcomes (e.g., survivability, transplant rejection, hypertension) with the physiological parameters of what the patient-specific model indicates. This work is currently in progress in collaboration with cardiologists to compare patient pathophysiology as predicted by the model versus their clinical outcomes and diagnoses.

4.7 Conclusion

In this chapter, I have described the process of selecting a published cardiovascular system model, parameterizing and optimizing it with hemodynamics data from right heart catheterization, and generating a patient-specific model. In theory, this process could be applied to a gamut of computational physiology model with parameters corresponding to available clinical data. For example, renal function models could be paired with blood electrolytes data. However, the process of selecting a suitable model and searching for a dataset that describes the corresponding properties and anatomies can be arduous without a more systematic approach. As such, in the next chapter, I describe an informatics pipeline that could better bridge the gap between clinical data and computational physiology models using semantic annotations.

Chapter 5. Linking Data to Models

As demonstrated in Chapter 4, computational physiology models can be made patient-specific using existing clinical data, and the benefits of patient-specific modeling with retrospective clinical data are described in Chapter 4, Section 4.6.1 and 4.6.2. While this approach for model optimization can be broadly applied, the *ad hoc* method for linking clinical data to model variables does not scale well as the number of data elements or model variables increase.

As I describe in Chapter 4, there is currently a large corpus of published models as discovered in the scoping review described in Chapter 3. At the same time, there is an abundance and a wide variety of clinical data available via the near-ubiquitous adoption of EHR systems. Albeit there are regulatory and accessibility barriers to using retrospective EHR data for research, there are existing and ongoing efforts to build and support research-ready clinical data repositories (Murphy et al., 2010). In order to take advantage of these resources, there needs to be a more precise and systematic method for linking clinical data to computational models. As such, I have developed an informatics pipeline for more systematically linking clinical data with computational models by leveraging semantic annotations.

5.1 Background and Motivation

As described in Chapter 2, biosimulation models are mathematical representations of biological processes. More specifically, computational physiology models represent physiological processes. Researchers often use such models to computationally test hypotheses about the mechanisms of underlying pathophysiology. Spurred by initiatives such as the Physiome Project and the Virtual Physiological Human (VPH) (Hunter and Viceconti, 2009), researchers are actively applying biosimulation modeling to advance personalized medical care and to improve drug design. These biosimulation models range in scale from subcellular processes, such as glycolysis (Vinnakota et

al., 2010) to that of fluid flow through the human circulatory system (Beard et al., 2013; Pettersen et al., 2014; Tewari et al., 2013, 2016). To validate and test these models, researchers match a model's simulated output to empirical physiological measurements. If the model can replicate the observed, measured behaviors, then it validates that the physiological theory is viable. This approach, from hypothesis, to animal validation, and to prospective human trials is well-proven, thorough, and systematic. However, it is very costly and is only ethical for data that can be collected via non-invasive, low-risk methods in human trials. In contrast, I propose that by using retrospective EHR data, researchers could validate physiological models in a more time- and cost-effective manner. As an example, this approach might allow researchers to more efficiently discard seemingly plausible hypotheses prior to the costly process of prospective human trials for validation.

Although only a small fraction of EHR data may be relevant for models of physiology, utilizing these data is still less expensive than designing a trial and prospectively collecting data from participants. By making these clinical data searchable and linkable to biosimulation models, researchers using biosimulation models could better find existing retrospective clinical data that facilitate the validation of hypotheses in humans. Conversely, clinical researchers can test hypotheses by finding appropriate models that are relevant to the clinical measurements of interest. Thus, a goal of this work is to connect biosimulation models to relevant EHR data for validation of those models, and likewise to connect EHR data to relevant biosimulation models. These connections could improve and allow for patient-specific modeling—the simulation of individual physiology and the potential to support clinical decisions in patient-specific treatment, prognosis, and diagnosis. In fact, there are ongoing patient-specific research efforts (Arthurs et al., 2016; Kim et al., 2010; Youssefi et al., 2018), and while these demonstrate important progress toward patient-

specific modeling, they require prospective collection of data specific to the models, which may not be practical nor part of the normal clinical workflow.

5.2 A Semantic Approach to Models and Data

My approach to connecting models to EHR data is semantics-based. I leverage existing biosimulation semantic annotation framework and ontologies to create mappings between clinical and computational physiology domains. The long-term goal is to develop a library of searchable biosimulation models that can be matched against physiological data measurements found in clinical data repositories. This library of annotated physiological models builds upon current libraries of models such as the BioModels Database (Le Novere et al., 2006) and the Physiome Model Repository (PMR) (Yu et al., 2011). In this chapter, I describe the informatics pipeline for annotating clinical data and matching these to annotated models. While this informatics pipeline is generalizable to any structured clinical data and computational physiology models, here, I demonstrate the pipeline with concrete examples using several quantitative clinical datasets available from the University of Washington Medical Center clinical data repository. Using clinical datasets from right heart catheterization and cardiac magnetic resonance imaging (cMRI), I describe how these two different datasets can be matched to the cardiovascular model used in Chapter 4 (*Minimal haemodynamic system model including ventricular interaction and valve dynamics* by Smith, et al.). In addition, I describe how a blood electrolytes lab panel dataset can be matched to two different renal function models and an ischemia model.

5.2.1 Semantics in Computational Physiology Models

Computational physiology and bioinformatics communities make good use of Semantic Web. For example, SBML models in BioModels Database and CellML models in the PMR both support semantic annotation. On the contrary, while the use of Semantic Web and domain ontologies in

healthcare is key for advanced analytics and automated reasoning, their use remains very limited (Atalag et al., 2017). In order to link these domains, it is imperative to represent and semantically annotate both computational and health information models using compatible standards and ontologies. Leveraging prior work in this area, I use biological and physical semantics as a common ground and annotate the data and model variables with composite annotations to unambiguously define both the physical property being measured, and the entity or process being involved (Gennari et al., 2011). The physical properties of interest may include attributes such as fluid volumes and pressures, flow rates, or chemical concentrations. To refer to these properties in an unambiguous manner, I use the Ontology of Physics for Biology (OBP), which contains a rich taxonomy of physical property types. For physical entities, I use knowledge resources such as Chemical Entities of Biological Interest (ChEBI) (Hastings et al., 2016) for chemicals, and the Foundational Model of Anatomy (FMA) for human anatomy. For example, "sodium ion concentration in blood" can be precisely described as:

> OPB:Chemical concentration <property_of> CHEBI:sodium(1+) <part_of> FMA:Portion of blood

The details of biomedical ontologies and the composite annotation framework are described in Chapter 2.

5.2.2 Clinical Informatics Standards and Ontologies

Clinical informatics standards and ontologies do in fact exist. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a very comprehensive controlled vocabulary encompassing more than 400,000 unique concepts and 1.4 million relationships. Its scope is the entire EHR and supports both care delivery and secondary use. Logical Observation Identifiers

Names and Codes (LOINC) is widely used for identifying laboratory tests, and it aims to establish common language and codes for identifying health measurements and observations.

However, there is still a gap between these health informatics standards and standards used by the biosimulation modeling community. This issue is partially addressed in a separate work in collaboration with K. Atalag, R. Kalbasi, and D. Nickerson, where we map clinical information standards to model standards using the openEHR information model, an information model for EHR data that specifies how health data should be represented, processed, and visualized.

5.3 A Pipeline for Semantic Annotation of Clinical Datasets and Models

Electronic health records are valuable sources of clinical data for research, but it takes significant effort to identify and transform raw EHR data into a form that can be used by researchers. In this section, I describe a bidirectional informatics pipeline for annotating clinical data with clinical standards and mapping them to model variable annotations.

As previously mentioned, I processed data from right heart catheterization, cardiac MRI, and blood electrolytes lab test datasets through the pipeline. All of these datasets contain quantitative measurements, making them amenable to connect with biosimulation models that simulate quantitative aspects of physiology.

5.3.1 Clinical Data

While this annotation methodology can be generalized to any clinical dataset, I will focus on quantitative clinical data from the De-identified Clinical Data Repository (DCDR) from University of Washington Institute of Translational Health Sciences. In recent years, initiatives like i2b2 (Informatics for Integrating Biology and the Bedside) have led to the development of clinical data

repositories, including the DCDR. Such repositories allow for much more efficient access to EHR data. In particular, the DCDR allows the researcher to directly query a de-identified subset of data from various UW Medicine clinical systems without requiring individual IRB approval, and without relying on a database administrator to query and relay the data. I obtained the blood electrolytes datasets from the DCDR by querying for "Labs > Chemistry > Blood Electrolytes". The entire blood electrolytes dataset includes 379,316 patients over a 5-year period, and I selected a random subset of 10,000 lab results.

In my initial approach, I transformed the source data from flat comma-separated value (CSV) format into CellML format using a Python script, which allowed me to annotate each data type using SemGen's model annotation tool as if it were a model variable. While this method made use of the available annotation tools for biosimulation models, it was not utilizing any existing clinical standards. In a more systematic approach outlined in Figure 5.1, the source data was imported into openEHR by mapping each data element to corresponding openEHR archetype nodes and annotating them with SNOMED or LOINC terms.

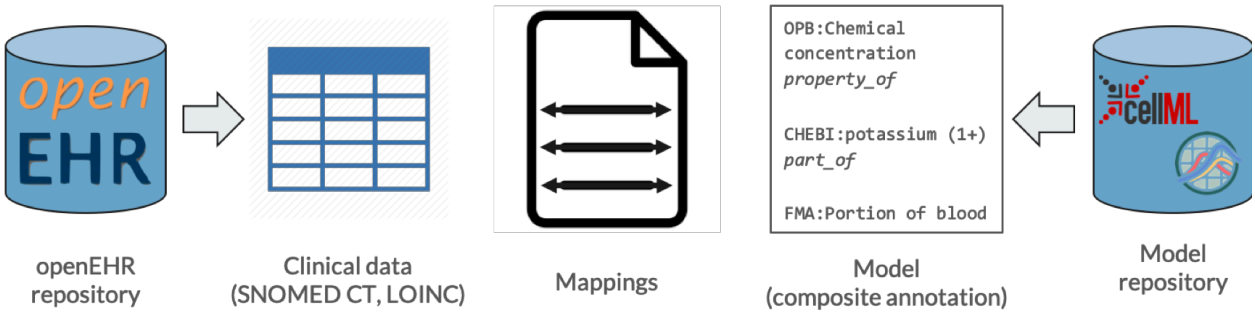

Figure 5.1 Bidirectional informatics pipeline for mapping EHR data to model variables.

5.3.2 Model Selection and Semantic Annotations

On the right-hand side of the pipeline in Figure 5.1 is the computational physiology model. The model can be selected from a model repository, such as the Physiome Model Repository or

BioModels. The selection can be driven either by the modeler with a specific model of interest, or by the clinical data of interest and the resulting model annotations mapped from those clinical annotations.

The selected model must have semantic annotations (i.e., composite annotations) of its variables in order to map these to appropriate data annotations. Recent work in the modeling community has encouraged this sort of semantic annotation and the use of the COMBINE Archive standard for archival storage of these annotations has gained traction.

5.4 Matching Annotated Datasets with Annotated Model Variables

Once the datasets were semantically annotated, I matched them with relevant biosimulation models. I matched data fields in the RHC and cardiac MRI datasets with variables from the Smith model (Table 1a and 1b). This cardiovascular systems model includes interactions between right and left ventricle and is able to evaluate the effects of pulmonary hypertension (increased right ventricular pressure) on cardiac function. Clinically, some of the RHC data are used to directly monitor and understand cardiac function: e.g., cardiac output as a measure of overall heart function. However, if I can use these data in conjunction with the Smith model, I can predict parameters in that mechanistic model that are not readily measurable or apparent in the RHC data. For example, the Smith model includes variables such as the elastance of the left ventricular free wall. Changes in such variables for a single patient over time have the potential to help clinicians better understand pathologies as they emerge, and better guide therapy for improved outcome.

Although non-invasive, cardiac MRIs include some of the same sorts of data; in particular, volumes of heart chambers can be readily measured, although pressures and flows cannot. Therefore, I matched the same Smith model to this dataset, albeit there were fewer matches to model variables. As with the RHC data, when I fit the model to the cardiac MRI dataset, changes

in model parameter values may correspond to specific diagnoses, such as cardiac hypertrophy, or heart failure with preserved ejection fraction (HFpEF).

Table 5.1 Examples of RHC data fields, their annotations, and matching model variables.

Data Field	Description	Composite Annotation	Model Variable (Source Model)
Pressure_AO	Aortic pressure	OPB:Fluid Pressure <property_of> FMA:Portion of blood <part_of> FMA:Aorta	P_ao (Smith)
Pressure_LV	Left ventricular pressure	OPB:Fluid Pressure <property_of> FMA:Portion of blood <part_of> FMA:Left ventricle	P_lv (Smith)
Pressure_RV	Right ventricular pressure	OPB:Fluid Pressure <property_of> FMA:Portion of blood <part_of> FMA:Right ventricle	P_rv (Smith)
Pressure_PA	Pulmonary artery pressure	OPB:Fluid Pressure <property_of> FMA:Portion of blood <part_of> FMA:Pulmonary artery	P_pa (Smith)
Thermo_Cardiac_Output	Cardiac output (thermodilution)	OPB:Fluid Flow Rate <property_of> Blood flow through aortic valve Source: Portion of blood in Left ventricle Sink: Portion of blood in Aorta	Q_av (Smith)
Pulmonary_Blood_Flow	Pulmonary blood flow	OPB:Fluid Flow Rate <property_of> Blood flow through pulmonary valve Source: Portion of blood in Right ventricle Sink: Portion of blood in Pulmonary artery	Q_pul (Smith)
Systemic_Blood_Flow	Systemic blood flow	OPB:Fluid Flow Rate <property_of> Systemic circulatory blood flow Source: Portion of blood in Aorta Sink: Portion of blood in Vena cava	Q_sys (Smith)
PulmVasc_Resistance	Pulmonary vascular resistance	OPB:Fluid Flow Resistance <property_of> FMA:Portion of blood	R_pul (Smith)

		<part_of> FMA:Pulmonary vascular system	
SV_Resistance	Systemic vascular resistance	OPB:Fluid Flow Resistance <property_of> FMA:Portion of blood <part_of> FMA:Systemic circulatory system	R_sys (Smith)
Heart_Rate	Heart rate	OPB:Temporal frequency	period (Smith)

Table 5.2 Examples of cardiac MRI data fields, their annotations, and matching model variables.

Data Field	Description	Composite Annotation	Model Variable (Source Model)
LVEDV	Left ventricular end diastolic volume	OPB:Fluid volume <property_of> FMA:Portion of blood <part_of> FMA:Left ventricle	V_lv (Smith)
LVESV	Left ventricular end systolic volume	OPB:Fluid volume <property_of> FMA:Portion of blood <part_of> FMA:Left ventricle	V_lv (Smith)
RVEDV	Right ventricular end diastolic volume	OPB:Fluid volume <property_of> FMA:Portion of blood <part_of> FMA:Right ventricle	V_rv (Smith)
RVESV	Right ventricular end systolic volume	OPB:Fluid volume <property_of> FMA:Portion of blood <part_of> FMA:Right ventricle	V_rv (Smith)
LVCO	Left ventricular cardiac output	OPB:Fluid Flow Rate <property_of> Blood flow through aortic valve Source: Portion of blood in Left ventricle Sink: Portion of blood in Aorta	Q_av (Smith)
RVCO	Right ventricular cardiac output	OPB:Fluid Flow Rate <property_of> Blood flow through pulmonary valve Source: Portion of blood in Right ventricle Sink: Portion of blood in Pulmonary artery	Q_pul (Smith)

Table 5.3 shows the same information for the third dataset, the blood electrolyte lab values, when compared to three different models. For the first two models, I focus on creatinine concentration, as both models use this data to calculate an estimate of the glomerular filtration rate of the kidney, an important indicator of kidney function. The first model by Waikar and Bonventre (Waikar and Bonventre, 2009) uses a two-compartment transport model to define the severity of acute kidney injury (AKI). The second model by Chen (Chen, 2013) calculates glomerular filtration rate (GFR) when the creatinine values change acutely over a 12 – 48 hour timespan, often as a result of AKI. Both models can be used to more accurately predict GFR during these rapid changes in blood serum creatinine, and the Waikar and Bonventre model can also be used to consider the effects of chronic kidney disease (CKD) if the patient has a previous diagnosis of CKD. The third model developed by Yi, et al. (Yi et al., 2003), tracks various ion concentrations during ischemia and hypoxia, and aims to determine the underlying mechanism behind increased extracellular potassium during ischemia.

Table 5.3 Examples of blood electrolytes data fields, their annotations, and matching model variables.

Data Field	Description	Composite Annotation	Model Variable (Source Model)
CRE - CREATININE	Concentration of creatinine in blood	OPB:Chemical concentration <property_of> CHEBI:creatinine <part_of> FMA:Portion of blood	C (Waikar); $Mean P_{Cr}$, DP_{Cr} (Chen)
NA - SODIUM	Concentration of sodium in blood	OPB:Chemical concentration <property_of> CHEBI:sodium(1+) <part_of> FMA:Portion of blood	Na_v (Yi)

K - POTASSIUM	Concentration of potassium in blood	OPB:Chemical concentration <property_of> CHEBI:potassium(1+) <part_of> FMA:Portion of blood	K_v (Yi)
CL - CHLORIDE	Concentration of chloride in blood	OPB:Chemical concentration <property_of> CHEBI:chloride(1-) <part_of> FMA:Portion of blood	Cl_v (Yi)

While Table 5.1, 5.2, and 5.3 show a number of semantic matches between data fields and model variables, the majority of EHR data does not fit the needs of the model. Some data fields, such as patient ID, are simply data that do not represent a physiological state or process. Others do represent interesting physiology but are not considered in the particular model of interest. For example, the left and right ventricular ejection fraction data from cardiac MRI, while very important indicators of heart function, does not have a semantic equivalence in the Smith model. Furthermore, some data may not have semantic equivalences in the model of interest, but they may be used to calculate another variable in the model. For example, I can use gender, height, and weight data in the RHC dataset with Nadler's formula to obtain an accurate estimate the patient's total blood volume, which does in fact have a corresponding variable in the Smith model.

Conversely, the model also contains numerous variables that do not match the EHR data. Some of these, such as "Gaussian Curve Delay," or "Pericardial Diastolic Exponential Parameter" are model parameters used to support the mathematics of model simulations. Others, such as "Pressure across the pericardial wall," or "Blood flow across tricuspid valve" are physiologically meaningful variables that are not available in the RHC or cardiac MRI dataset, and very unlikely to be captured in a clinical setting.

In addition, to better connect data to models, the researcher should consider contextual information. For example, the Waikar model is for pathophysiology—physiology under a diseased condition. Thus, it would be more appropriate to validate this model with EHR lab data that is collected only

from patients in that diseased state. In the openEHR model, I can easily create a selection query over my defined archetypes that retrieves only those lab values with a certain semantic constraint matching the desired SNOMED term—namely, for the Waikar model, those lab results that include a diagnosis of "kidney disease" per the constraint on a creatinine test.

5.5 Summary

I have described an informatics pipeline for matching EHR datasets with biosimulation models that serves as a modest first step towards connecting clinical data with physiological models. Although portions of this pipeline are manually created, some of these manual steps could be replaced by automation. Although for models with a limited number of variables, manual mappings may be sufficient. In the long term, mapping clinical data to biosimulation models has the potential to benefit both biosimulation modelers by better facilitating the use of retrospective clinical data for model validation, and clinical researchers by better connecting biosimulation models that can simulate patient physiology using available clinical data.

I am streamlining the process for physiological modelers to use EHR data as a new valuable resource for model validation. Model validation is currently a time-consuming and costly step in creating biosimulation models. Simultaneously, I am streamlining the process for clinical researchers to use biosimulation models. While the EHR data may not fit the needs of the modeler as well as a prospectively designed data collection experiment, it does provide a much easier avenue for collecting data with sample sizes that are often infeasible in a prospective experiment.

Additionally, connecting EHR data to biosimulation models has the potential to make biosimulation models patient-specific. By parameterizing a model with a patient's data, the given models could better track the progression of acute kidney injury or ionic imbalance during ischemia at an individual level. Furthermore, by optimizing the parameters not explicitly defined

in the model to simulate the patient's data, the model could be used to estimate values that describe underlying physiological features that are not readily measurable in a clinical setting. It is important to note that not all clinical data are going to be relevant to the model. This is especially true for unstructured or less quantitative data, such as medical history and clinical notes. While these are all valuable information for the assessment and treatment of the patient, computational models cannot directly utilize these data.

In conclusion, I have described an informatics pipeline for applying clinical standards to EHR datasets and mapping them with computational physiology model annotations. I have also demonstrated how this pipeline can connect specific EHR data with specific biosimulation models, using blood electrolytes datasets as an example. My work aims to support clinical research informatics; by connecting physiological models to EHR data, I can better support the basic research being carried out by physiologists, with the long-term potential of developing patient-specific models for improved clinical decision making.

Chapter 6. Conclusion

In this chapter, I conclude my dissertation by summarizing the research, identifying its limitations, and discussing broader implications for biosimulation modelers and clinicians. Furthermore, I explore future directions stemming from this work beyond the scope of this dissertation.

6.1 Dissertation Summary

In this dissertation, I have described my work using retrospective EHR data to optimize and validate a biosimulation model. By optimizing a cardiovascular systems model with patient-specific data, I have demonstrated the feasibility of patient-specific modeling using existing clinical data and a biosimulation model available from a model repository.

In Chapter 2, I provide the necessary background information on biosimulation modeling, and EHR data. Specifically, I describe the standards used in both fields, as well as annotation efforts in the modeling community.

In Chapter 3, I describe the scoping review of biosimulation models in literature. Motivated by the push from the modeling community towards reproducible models, I reviewed 150 biosimulation model publications, consisting of 50 general biosimulation models from PubMed, 50 cardiovascular biosimulation models from PubMed, and 50 diabetes models from a review article by Ajmera, et al. From the 150 model publications, I discovered a stark lack of reproducibility as characterized by just 2 model publications making the model source code available.

In Chapter 4, I describe model optimization using retrospective EHR data. Using right heart catheter hemodynamics dataset from UW Medicine Regional Heart Center, I optimize a subset of parameters in a cardiovascular model from the Physiome Model Repository to fit the RHC data. By optimizing this model with patient-specific data, I demonstrate patient-specific modeling using

existing clinical data. Furthermore, I discuss the potential uses for patient-specific modeling for clinicians, including estimating values that are physiologically interesting but clinically difficult to measure, and generating PV loops from right heart catheterization alone.

In Chapter 5, I describe my approach to generalizing the connection between clinical data and biosimulation models. In my approach, I leverage biomedical ontologies and the composite annotation framework, and map the model annotations to clinical controlled terminologies.

6.2 Broader Implications

While I mainly focused on a cardiovascular example for model validation and patient-specific modeling, the same methodology can be extended to other clinical and biological domains. Especially as more models are semantically annotated, and more data-model mappings are created and stored, semantic queries will allow for more automatic and precise search for models with corresponding data, and vice versa.

The broader implications of this work can be divided into two main categories: Implications for biosimulation modelers; and implications for clinicians. This two-pronged impact bridges the gap in translational research to benefit not only clinical practice, but also to benefit and accelerate basic research by leveraging the ever-growing volume of clinical data.

6.2.1 Modeling Implications

For biosimulation modelers, this work demonstrates the feasibility of using retrospective EHR data for model validation. As a part of the model development process, the modeler must validate his or her model against data. Typically, the modeler must conduct a prospective study to collect the necessary data. Whether the data is collected from human subjects, or through laboratory benchwork experiments, the data collection process can be time-consuming, and costly.

Furthermore, data collected from human subjects can be limited by the sample size, especially for a subject cohort with very specific inclusion criteria. The data type collected is also limited to those that can only be collected by procedures that do not harm the human subject.

My approach of using retrospective data can potentially alleviate the burden of arduous data collection required for model validation. With retrospective data, the data has already been collected through ordinary course of care. The only burden for the modeler is finding the dataset that corresponds to the model and optimizing the model to estimate selected model parameters not directly prescribed by the corresponding data. In addition to the retrospective nature of data alleviating the burden of data collection, my approach utilizes clinical data collected from actual patients. Especially for modelers studying human diseases, human pathophysiological data cannot be readily substituted by experimental animal data or canonical values. While this is can be a non-trivial process, the model optimization approach in Chapter 4 and the semantic mapping approach in Chapter 5 addresses these issues. By providing modelers with an alternative method for acquiring data without burdensome prospective data collection, my approach has the potential to lower the barrier and accelerate biosimulation model development.

6.2.2 Clinical Implications

For clinicians, patient-specific models can serve as a powerful clinical decision support tool. Using patient data ordinarily collected through the course of care, my approach enables those data to be coupled with biosimulation models that represent the relevant physiology. The biosimulation models can then be optimized to patient data, making the model patient-specific, and representative of the patient's pathophysiology.

The patient-specific model has several benefits for the clinicians. First, it can paint a more holistic picture of the patient's physiology. The work described in Chapter 4 with the Smith model and

right heart catheter hemodynamics data shows how blood pressure measurements can be used to estimate volume, resistance, and elastance measurements. These additional estimated values provide the clinician with additional information regarding the patient's physiology that could help clinicians make more informed clinical decisions. For example, using patient-specific hemodynamics data with a cardiovascular model, the model can estimate physiologically interesting values like the elastance of pulmonary veins, which cannot be directly measured. Additional physiological information like this could be useful in elucidating the pathophysiology of poorly understood diseases like pulmonary hypertension, where different sub-categories of pulmonary hypertension have vastly different treatment options, and identifying the root cause is crucial but difficult.

Because the model can estimate volume measurements from the pressure measurements, and because it estimates the time course measurements for pressure and volume, the model is able to generate patient-specific pressure-volume (PV) loops, which plots the pressure against the volume at corresponding time points. The left ventricular PV loop provides a framework for clinicians and clinical researchers to quickly understand the cardiac function, including stroke volume, cardiac output, ejection fraction, as well as determine various cardiac abnormalities based on the shape of the PV loop. While the PV loop is a useful tool, it requires simultaneous capture of pressure and volume in the cardiac system. However, hemodynamics measurements are typically limited to capturing one physical property at a time: Cardiac catheterization captures pressures, but not volumes; Echocardiography captures volumes, but not pressures. As a result, PV loops are very difficult to generate in a clinical setting without specialized equipment. However, as demonstrated in Chapter 4, they can be estimated from pressure measurements using biosimulation models.

Given longitudinal patient data, the model could also be used to track patient trajectory over time to monitor how the patient’s pathophysiology changes over a period of time. As stated in Section

4.6.1, this type of longitudinal analysis is performed in a collaboration with Woodall, et al. to track cardiac function in heart transplant patients following the transplantation (Woodall et al., 2018). In this work, we used longitudinal right heart catheter hemodynamics data from heart transplant patients to create patient-specific models at various time points following the heart transplant. With this analysis, we estimated various cardiac functions over time, including ejection fraction, aortic elastance, and systemic resistance. Based on these cardiac functions following the transplant, we were able to see distinct progressions of cardiac function for different patients.

Furthermore, the model could be used to conduct perturbation studies. For example, parameter values for the elastance of aorta and the arterial vasculature could be increased to explore the downstream effects of arterial stiffening on the rest of the patient's hemodynamics system.

6.3 Research Limitations

While my work demonstrates the feasibility of using a published model with retrospective EHR data for model validation and patient-specific modeling, it has limitations that could be addressed to bolster the research work and move towards widespread and practical application. Overall, this dissertation is a step towards enabling the secondary use of EHR data for biosimulation modeling. However, it is limited in scope, lacks rigorous clinical validation, and relies on manual steps.

As described in section 3.5.1, my scoping review was limited only to publications available in PubMed. While PubMed contains publications in life sciences and biomedicine, there are other journals not index in PubMed that contain biosimulation model publications, and relevant for this scoping review. Some of these journals not included in PubMed pertain to mathematics and engineering, and these journals may have more stringent requirement on publishing the source code or data. The overall sample size of publications analyzed in the scoping review was also limited. Nonetheless, this work revealed a consistent lack of the source code in model publications

I reviewed, and more publications from a wider range of journals could be reviewed to address this weakness.

Perhaps the biggest limitation of my dissertation is the lack of rigorous clinical validation of the patient-specific models in Chapter 4. The patient-specific models were mathematically valid, and produced outputs with reasonable residual values compared to the patient data. The models also elucidated information about the patient's physiology. For patient-specific models at single time point, the results in Section 4.5 indicates systolic dysfunction in the right heart for patient 266 and 572. Furthermore, the longitudinal analysis described in Section 4.6.1 and Woodall, et al., predicted patients with decreasing cardiac function over time, suggesting a less successful recovery from the heart transplant. However, this model-derived information was not clinically validated, and it is unclear if the model correctly estimated the pathophysiology of the patient. Conducting a chart review, or prospective study to track the patient outcomes is needed to confirm the validity of the model predictions. In addition, for patient-specific modeling to be practical, especially at the point of care in a clinical setting, the runtime for the optimization would need to be reduced drastically. With my optimizations, the runtime ranged in the scale of hours to days. Reducing the model complexity could reduce the optimization time. Perhaps if patient-specific modeling were to be used in clinical setting, the model could be developed specifically with the available data in mind. With a model that simply used available data as inputs, and produced useful output, time-consuming optimizations would be unnecessary.

Finally, the mapping of clinical terminologies with model annotations described in Chapter 5 is certainly a more systematic approach than an *ad hoc* approach to linking clinical data with biosimulation models, leveraging precise semantics that make the linkages more reusable. However, the current approach still relies on creating manual mappings. Similar limitation exists for model annotation, where the annotation of model variables still remains a manual process in

which the annotator must clearly understand the model and its variables. Although this process may require a large initial investment in creating the mappings and annotations, in the long run it will enable better interoperability between clinical data and models.

6.4 Future Directions

There are several directions for future research that can stem from this work. In my scoping review of model literature described in Chapter 3, I discovered and quantified the lack of model source code availability from publications. While this work did not identify the root cause, it could be used to drive broader change in the modeling community, and perhaps more importantly for journals and publishers to encourage, if not require, authors to publish their data and source code. In order for computational experiments to be reproducible, the publication must include all of the necessary information to repeat the experiment. There are efforts, such as the Center for Reproducible Biomedical Modeling, that are working with journals to annotate the models and make them more reusable. My scoping review could be synergistic with such efforts in moving the broader modeling community and journals towards more reproducible models. Using my scoping review as a baseline, another interesting and useful study would be to stratify the mode code availability by year to identify any trends in model publication. Especially if such study were to be done several years after the establishment of the Center for Reproducible Biomedical Modeling, it could be reveal the potential effect the center might have on the publication culture and reproducibility of biosimulation models in literature.

In Chapter 4, I described patient-specific modeling of the cardiovascular system using RHC datasets. The patient-specific model in this work is able to extrapolate volumetric data by optimizing the model parameters to the pressure measurements taken through catheterization process, which involves making an incision and inserting a catheter through the patient's superior

vena cava into the right side of the heart. While this procedure is performed routinely for heart transplant patients, it is still an invasive procedure. A future work for this work would be to apply the same model optimization and patient-specific modeling approach to a cardiovascular model with data from less invasive procedures, such as cardiac MRI or echocardiograms. If patient-specific modeling could be used to extrapolate pressure values, typically measured invasively, from noninvasive volumetric data, it could be a valuable tool for clinicians and patients for replacing invasive procedures with noninvasive procedures. Especially with additional future work with clinical validation of the results of Chapter 4, this approach of using non-invasive measurements for patient-specific modeling could have tremendous clinical impact.

The data-to-model mapping work in Chapter 5 could be extended to connect more clinical data and model annotations. This work could be accelerated by advancements in the field of ontology alignment. Not only could the number of mappings be increased, but these mappings could make use of existing annotation archiving standards and potential future annotation infrastructure. In order to make the mappings more accessible and usable, they should be archived in a standard structure. One promising near-term potential would be to store these mappings, along with the model file and the data file, in a COMBINE archive. COMBINE archive is already in use by the modeling community as a method of bundling the model file with the supporting files, such as model annotation and experimental data (Bergmann et al., 2014; Neal et al., 2018a). For a long-term future work, the mappings could be stored in a centralized knowledgebase, where the community can upload new mappings between clinical concepts and model annotations, or access previously created mappings, so as to not expend efforts in recreating mappings that others have already created.

The grand vision of this work is as follows: A clinician would like to better understand the pathophysiology of a patient. He or she queries a biosimulation model knowledge base and finds

models with semantic overlaps with the patient data. Using this model, the clinician performs patient-specific modeling at the point of care to better understand the patient's physiology, and to conduct perturbation studies to simulate how the patient's physiology is affected by various treatment options. Especially with better interoperability standards like FHIR and CDS hooks, patient-specific modeling could be a powerful tool as a part of the clinical workflow.

6.5 Final Conclusion

In this dissertation, I have described my work on the secondary usage of electronic health record data for patient-specific modeling. I conducted a scoping review to better understand the characteristics and reproducibility of biosimulation models in literature. By optimizing a published cardiovascular model to patient-specific right heart catheter hemodynamics data, I have demonstrated the feasibility of patient-specific modeling using retrospective clinical data. Extending the process of matching data to model parameters, I have described an annotation pipeline and a mapping approach to systematically link clinical data to biosimulation models.

My dissertation makes use of existing resources, including retrospective EHR data and published models, to create additional value for modelers and clinicians. My work has the potential to accelerate biosimulation modeling by paving the way to a new approach to validate models without the burden of prospective data collection. Furthermore, the patient-specific modeling described in my dissertation can help clinicians better understand the patient’s pathophysiology and make more informed clinical decisions. Patient-specific models could be used to estimate additional physiological values that cannot be measured otherwise in a clinical setting, or to conduct perturbation studies that simulate the patient physiology. Especially for diseases that are difficult to diagnose, such as pulmonary hypertension or heart failure with preserved ejection fraction (HFpEF), patient-specific modeling could be a valuable clinical tool.

BIBLIOGRAPHY

Adler-Milstein, J., Holmgren, A.J., Kralovec, P., Worzala, C., Searcy, T., and Patel, V. (2017). Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J. Am. Med. Inform. Assoc. JAMIA *24*, 1142–1148.

Ajmera, I., Swat, M., Laibe, C., Le Novère, N., and Chelliah, V. (2013). The impact of mathematical modeling on the understanding of diabetes and related complications. CPT Pharmacomet. Syst. Pharmacol. *2*, e54.

Amin, A., Taghavi, S., Esmaeilzadeh, M., Bakhshandeh, H., Naderi, N., and Maleki, M. (2011). Pulmonary Arterial Elastance for Estimating Right Ventricular Afterload in Systolic Heart Failure: right ventricular afterload in systolic heart failure. Congest. Heart Fail. *17*, 288–293.

Anderson, S., Allen, P., Peckham, S., and Goodwin, N. (2008). Asking the right questions: Scoping studies in the commissioning of research on the organisation and delivery of health services. Health Res. Policy Syst. *6*, 7.

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. Int. J. Soc. Res. Methodol. *8*, 19–32.

Arthurs, C.J., Lau, K.D., Asrress, K.N., Redwood, S.R., and Figueroa, C.A. (2016). A Mathematical Model of Coronary Blood Flow Control: Simulation of Patient-Specific Three-Dimensional Hemodynamics during Exercise. Am. J. Physiol. - Heart Circ. Physiol. ajpheart.00517.2015.

Atalag, K., Kalbasi, R., and Nickerson, D. (2017). A Semantic Web based Framework for Linking Healthcare Information with Computational Physiology Models. In HiNZ 2017 Conference Proceedings, (Rotorua, New Zealand), p.

Bassingthwaighte, J. (2000). Strategies for the Physiome Project. Ann. Biomed. Eng. *28*, 1043–1058.

Beard, D.A., Pettersen, K.H., Carlson, B.E., Omholt, S.W., and Bugenhagen, S.M. (2013). A computational analysis of the long-term regulation of arterial pressure. F1000Research.

Bergmann, F.T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., Hucka, M., Laibe, C., Miller, A.K., Nickerson, D.P., et al. (2014). COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. BMC Bioinformatics *15*, 369.

Butterworth, E., Jardine, B.E., Raymond, G.M., Neal, M.L., and Bassingthwaighte, J.B. (2014). JSim, an open-source modeling system for data analysis. F1000Research.

Caroli, A., Manini, S., Antiga, L., Passera, K., Ene-Iordache, B., Rota, S., Remuzzi, G., Bode, A., Leermakers, J., van de Vosse, F.N., et al. (2013). Validation of a patient-specific hemodynamic computational model for surgical planning of vascular access in hemodialysis patients. Kidney Int. *84*, 1237–1245.

Chen, S. (2013). Retooling the creatinine clearance equation to estimate kinetic GFR when the plasma creatinine is changing acutely. J. Am. Soc. Nephrol. JASN *24*, 877–888.

Cook, D.L., Mejino, J.L.V., Neal, M.L., and Gennari, J.H. (2008). Bridging Biological Ontologies and Biosimulation: The Ontology of Physics for Biology. AMIA. Annu. Symp. Proc. *2008*, 136–140.

Cook, D.L., Bookstein, F.L., and Gennari, J.H. (2011). Physical properties of biological entities: an introduction to the ontology of physics for biology. PLoS One *6*, e28708.

Cuellar, A.A., Lloyd, C.M., Nielsen, P.F., Bullivant, D.P., Nickerson, D.P., and Hunter, P.J. (2003). An Overview of CellML 1.1, a Biological Model Description Language. SIMULATION *79*, 740–747.

Davis, K., Drey, N., and Gould, D. (2009). What are scoping studies? A review of the nursing literature. Int. J. Nurs. Stud. *46*, 1386–1400.

FHIR Specification HL7 Standards Product Brief - HL7 Fast Healthcare Interoperability Resources Specification (FHIR®), DSTU Release 1.

Forrey, A.W., McDonald, C.J., DeMoor, G., Huff, S.M., Leavelle, D., Leland, D., Fiers, T., Charles, L., Griffin, B., Stalling, F., et al. (1996). Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin. Chem. *42*, 81–90.

Gennari, J.H., Neal, M.L., Galdzicki, M., and Cook, D.L. (2011). Multiple ontologies in action: Composite annotations for biosimulation models. J. Biomed. Inform. *44*, 146–154.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. *44*, D1214-9.

Hersh, W.R. (2007). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am. J. Manag. Care *13*, 277–278.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinforma. Oxf. Engl. *19*, 524–531.

Hunter, P.J., and Borg, T.K. (2003). Integration from proteins to organs: the Physiome Project. Nat. Rev. Mol. Cell Biol. *4*, 237–243.

Hunter, P.J., and Viceconti, M. (2009). The VPH-Physiome Project: Standards and Tools for Multiscale Modeling in Clinical Applications. IEEE Rev. Biomed. Eng. *2*, 40–53.

Hunter, P., Robbins, P., and Noble, D. (2002). The IUPS human physiome project. Pflüg. Arch. *445*, 1–9.

IDC (2014). The Digital Universe Driving Data Growth in Healthcare (EMC Digital Universe).

Kalra, D., Beale, T., and Heard, S. (2005). The openEHR Foundation. Stud. Health Technol. Inform. *115*, 153–173.

Kim, H.J., Vignon-Clementel, I.E., Coogan, J.S., Figueroa, C.A., Jansen, K.E., and Taylor, C.A. (2010). Patient-Specific Modeling of Blood Flow and Pressure in Human Coronary Arteries. Ann. Biomed. Eng. *38*, 3195–3209.

Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., et al. (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. *34*, D689–D691.

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M.I., et al. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst. Biol. *4*, 92–92.

McGuinness, D.L., Van Harmelen, F., and others (2004). OWL web ontology language overview. W3C Recomm. *10*, 2004.

Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. *17*, 124–130.

Nadler, S.B., Hidalgo, J.U., and Bloch, T. (1962). Prediction of blood volume in normal human adults. Surgery *51*, 224–232.

Neal, M.L., and Kerckhoffs, R. (2009). Current progress in patient-specific modeling. Brief. Bioinform. bbp049.

Neal, M.L., Gennari, J.H., Arts, T., and Cook, D.L. (2009). Advances in Semantic Representation for Multiscale Biosimulation: A Case Study in Merging Models. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 304–315.

Neal, M.L., Carlson, B.E., Thompson, C.T., James, R.C., Kim, K.G., Tran, K., Crampin, E.J., Cook, D.L., and Gennari, J.H. (2015). Semantics-Based Composition of Integrated Cardiomyocyte Models Motivated by Real-World Use Cases. PloS One *10*, e0145621.

Neal, M.L., König, M., Nickerson, D., Misirli, G., Kalbasi, R., Dräger, A., Atalag, K., Chelliah, V., Cooling, M.T., Cook, D.L., et al. (2018a). Harmonizing semantic annotations for computational models in biology. Brief. Bioinform.

Neal, M.L., Thompson, C.T., Kim, K.G., James, R.C., Cook, D.L., Carlson, B.E., and Gennari, J.H. (2018b). SemGen: a tool for semantics-based annotation and composition of biosimulation models. Bioinforma. Oxf. Engl.

Nickerson, D.P., and Hunter, P.J. (2017). Introducing the Physiome Journal: Improving Reproducibility, Reuse, and Discovery of Computational Models. In 2017 IEEE 13th International Conference on E-Science (e-Science), pp. 448–449.

Ohno-Machado, L. (2014). NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. J. Am. Med. Inform. Assoc. JAMIA *21*, 193.

Pak, H. Unstructured Data In Healthcare.

Patil, R.K., Goyal, P., Swaminathan, R.V., Kim, L.K., and Feldman, D.N. (2017). Invasive Hemodynamic Assessment of Patients with Heart Failure and Pulmonary Hypertension. Curr. Treat. Options Cardiovasc. Med. *19*, 40.

Penicka, M., Bartunek, J., Trakalova, H., Hrabakova, H., Maruskova, M., Karasek, J., and Kocka, V. (2010). Heart failure with preserved ejection fraction in outpatients with unexplained dyspnea: a pressure-volume loop analysis. J. Am. Coll. Cardiol. *55*, 1701–1710.

Pettersen, K.H., Bugenhagen, S.M., Nauman, J., Beard, D.A., and Omholt, S.W. (2014). Arterial stiffening provides sufficient explanation for primary hypertension. PLoS Comput. Biol. *10*, e1003634.

Reis, Z.S.N., Maia, T.A., Marcolino, M.S., Becerra-Posada, F., Novillo-Ortiz, D., and Ribeiro, A.L.P. (2017). Is There Evidence of Cost Benefits of Electronic Medical Records, Standards, or Interoperability in Hospital Information Systems? Overview of Systematic Reviews. JMIR Med. Inform. *5*, e26.

Rosse, C., and Mejino, J. (2008). The Foundational Model of Anatomy Ontology. In Anatomy Ontologies for Bioinformatics, A.B.Bs. MSc, D.D. BSc, and R.B. BSc, eds. (Springer London), pp. 59–117.

Rosse, C., and Mejino, J.L.V. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J. Biomed. Inform. *36*, 478–500.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. ArXiv160904747 Cs.

Sauro, H.M., Blinov, M., Cook, D., Gennari, J.H., Goldberg, A., Karr, J., Moraru, I., Nickerson, D., and Schaff, J.C. (2018). Center for Reproducible Biomedical Modeling. (Honolulu, HI), p.

Sivanandam, S.N., and Deepa, S.N. (2008). Genetic Algorithm Optimization Problems. In Introduction to Genetic Algorithms, S.N. Sivanandam, and S.N. Deepa, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 165–209.

Smith, B.W., Chase, J.G., Nokes, R.I., Shaw, G.M., and Wake, G. (2004). Minimal haemodynamic system model including ventricular interaction and valve dynamics. Med. Eng. Phys. *26*, 131–139.

SNOMED International SNOMED Clinical Terms (SNOMED CT).

Tewari, S.G., Bugenhagen, S.M., Wang, Z., Schreier, D.A., Carlson, B.E., Chesler, N.C., and Beard, D.A. (2013). Analysis of cardiovascular dynamics in pulmonary hypertensive C57BL6/J mice. Front. Physiol. *4*.

Tewari, S.G., Bugenhagen, S.M., Vinnakota, K.C., Rice, J.J., Janssen, P.M.L., and Beard, D.A. (2016). Influence of metabolic dysfunction on cardiac mechanics in decompensated hypertrophy and heart failure. J. Mol. Cell. Cardiol. *94*, 162–175.

Townsend, R.R. (2017). Arterial Stiffness: Recommendations and Standardization. Pulse *4*, 3–7.

Vinnakota, K.C., Rusk, J., Palmer, L., Shankland, E., and Kushmerick, M.J. (2010). Common phenotype of resting mouse extensor digitorum longus and soleus muscles: equal ATPase and glycolytic flux during transient anoxia. J. Physiol. *588*, 1961–1983.

Waikar, S.S., and Bonventre, J.V. (2009). Creatinine kinetics and the definition of acute kidney injury. J. Am. Soc. Nephrol. JASN *20*, 672–679.

Woodall, N.P., Kim, K.G., Colunga, A.L., Gennari, J.H., Olufsen, M.S., and Carlson, B.E. (2018). Deep phenotyping of cardiac function in heart transplant patients using cardiovascular systems models. Manuscript submitted for publication to J. Appl. Physiol.

Yi, C.S., Fogelson, A.L., Keener, J.P., and Peskin, C.S. (2003). A mathematical study of volume shifts and ionic concentration changes during ischemia and hypoxia. J. Theor. Biol. *220*, 83–106.

Youssefi, P., Gomez, A., Arthurs, C., Sharma, R., Jahangiri, M., and Alberto Figueroa, C. (2018). Impact of Patient-Specific Inflow Velocity Profile on Hemodynamics of the Thoracic Aorta. J. Biomech. Eng. *140*.

Yu, T., Lloyd, C.M., Nickerson, D.P., Cooling, M.T., Miller, A.K., Garny, A., Terkildsen, J.R., Lawson, J., Britten, R.D., Hunter, P.J., et al. (2011). The Physiome Model Repository 2. Bioinformatics *27*, 743–744.

VITA

Graham Karam Kim was born in Seoul, South Korea and grew up in Portland, Oregon. He earned his Bachelor of Science in Bio and Brain Engineering from Korea Advanced Institute of Science and Technology (KAIST). In 2013, he moved to Seattle, Washington to pursue his interests in secondary usage of biomedical data. He earned his Doctor of Philosophy in Biomedical and Health Informatics from the University of Washington in 2018.