



University of Pennsylvania
ScholarlyCommons


Publicly Accessible Penn Dissertations

2021

Computer-Aided Clinical Trials For Medical Devices

Kuk Jin Jang
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Computer Sciences Commons](#), and the [Electrical and Electronics Commons](#)

Recommended Citation

Jang, Kuk Jin, "Computer-Aided Clinical Trials For Medical Devices" (2021). *Publicly Accessible Penn Dissertations*. 4776.
<https://repository.upenn.edu/edissertations/4776>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/4776>
For more information, please contact repository@pobox.upenn.edu.

Computer-Aided Clinical Trials For Medical Devices

Abstract

Life-critical medical devices require robust safety and efficacy to treat patient populations with potentially large patient heterogeneity. Today, the de facto standard for evaluating medical devices is the randomized controlled trial. However, even after years of device development many clinical trials fail. For example, in the Rhythm ID Goes Head to Head Trial (RIGHT) the risk for inappropriate therapy by implantable cardioverter defibrillators (ICDs) actually increased relative to control treatments. With recent advances in physiological modeling and devices incorporating more complex software components, population-level device outcomes can be obtained with scalable simulations. Consequently, there is a need for data-driven approaches to provide early insight prior to the trial, lowering the cost of trials using patient and device models, and quantifying the robustness of the outcome.

This work presents a clinical trial modeling and statistical framework which utilizes simulation to improve the evaluation of medical device software, such as the algorithms in ICDs. First, a method for generating virtual cohorts using a physiological simulator is introduced. Next, we present our framework which combines virtual cohorts with real data to evaluate the efficacy and allows quantifying the uncertainty due to the use of simulation. Results predicting the outcome of RIGHT and improving statistical power while reducing the sample size are shown. Finally, we improve device performance with an approach using Bayesian optimization. Device performance can degrade when deployed to a general population despite success in clinical trials. Our approach improves the performance of the device with outcomes aligned with the MADIT-RIT clinical trial. This work provides a rigorous approach towards improving the development and evaluation of medical treatments.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Electrical & Systems Engineering

First Advisor

Rahul Mangharam

Keywords

Applied Bayesian methods, Clinical trials, Computer modeling and simulation, Generative models, Medical devices, Uncertainty quantification

Subject Categories

Computer Sciences | Electrical and Electronics

COMPUTER-AIDED CLINICAL TRIALS FOR MEDICAL DEVICES

Kuk Jin Jang

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Rahul Mangharam, PhD, Associate Professor of Electrical and Systems Engineering,
University of Pennsylvania

Graduate Group Chairperson

Alejandro Ribeiro, PhD, Professor of Electrical and Systems Engineering,
University of Pennsylvania

Dissertation Committee:

Rajeev Alur, PhD, Zisman Family Professor of Computer and Information Science,
University of Pennsylvania

Pratik Chaudhari, PhD, Assistant Professor of Electrical and Systems Engineering,
University of Pennsylvania

James Weimer, PhD, Research Assistant Professor of Computer and Information Science,
University of Pennsylvania

Jin-oh Hahn, PhD, Associate Professor of Mechanical Engineering,
University of Maryland

COMPUTER-AIDED CLINICAL TRIALS FOR MEDICAL DEVICES

© COPYRIGHT

2021

Kuk Jin Jang

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 4.0
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

Dedicated to God, my beloved parents & family

ACKNOWLEDGEMENT

First and foremost, I would like to thank Professor Rahul Mangaharam and my thesis committee for their guidance and supervision during the completion of this thesis. I thank Professor Mangharam for giving me this opportunity and supporting me throughout my PhD and Professor James Weimer for his advice and encouragement.

Next, throughout the years, I've had the great chance to meet many amazing individuals and collaborate with them. I would like to express my gratitude to the my collaborators who have helped me in my research: Houssam Abbas, Zhihao Jiang, Yash Vardhan Pant, Marco Beccani, Jackson Liang, Sanjay Dixit, Bo Zhang, Marco Becanni, Renukanandan Tumu, Nicole Chiou, Ali Tivay, Sam Huang, Madeline Diep, Prof. Eric Eaton, Prof. Pratik Chaudhari and more.

I thank my friends and family who have cheered me on with never-ending support: My mother and father, Donghun Lee, Hyungwon Kim, Tasos Tsiamos, Alena Rodionova, Achin Jain, Heejin Jeong, Matei Ionita, Andreea Alexandru, Seungjoo Kum, Jinsung Kim, Sooyong Jang, all my friends and colleagues at the Graduate Student Center and many more. I'd like to express special thanks to Jo Ellen McBride for all the advice as well as all the efforts in helping me revise this thesis.

I thank all my labmates who helped me achieve this accomplishment: Billy Hongrui Zheng, Nandan Tumu, Johannes Betz, and Jiyue He. Their support was pivotal in helping me complete this process. Finally, I would like to extend a special thanks to Matthew O'Kelly. He has been an amazing friend and colleague and his help was essential for the completion of this thesis.

ABSTRACT

COMPUTER-AIDED CLINICAL TRIALS FOR MEDICAL DEVICES

Kuk Jin Jang

Rahul Mangharam

Life-critical medical devices require robust safety and efficacy to treat patient populations with potentially large patient heterogeneity. Today, the de facto standard for evaluating medical devices is the randomized controlled trial. However, even after years of device development many clinical trials fail. For example, in the Rhythm ID Goes Head to Head Trial (RIGHT) the risk for inappropriate therapy by implantable cardioverter defibrillators (ICDs) actually increased relative to control treatments. With recent advances in physiological modeling and devices incorporating more complex software components, population-level device outcomes can be obtained with scalable simulations. Consequently, there is a need for data-driven approaches to provide early insight prior to the trial, lowering the cost of trials using patient and device models, and quantifying the robustness of the outcome.

This work presents a clinical trial modeling and statistical framework which utilizes simulation to improve the evaluation of medical device software, such as the algorithms in ICDs. First, a method for generating virtual cohorts using a physiological simulator is introduced. Next, we present our framework which combines virtual cohorts with real data to evaluate the efficacy and allows quantifying the uncertainty due to the use of simulation. Results predicting the outcome of RIGHT and improving statistical power while reducing the sample size are shown. Finally, we improve device performance with an approach using Bayesian optimization. Device performance can degrade when deployed to a general population despite success in clinical trials. Our approach improves the performance of the device with outcomes aligned with the MADIT-RIT clinical trial. This work provides a rigorous approach towards improving the development and evaluation of medical treatments.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iv
ABSTRACT	v
LIST OF TABLES	ix
LIST OF ILLUSTRATIONS	xii
CHAPTER 1 : Introduction	1
1.1 Clinical Trials for Medical Treatment Evaluation	1
1.2 Limitations of Clinical Trials for Evaluating Life-critical Medical Devices . .	2
1.3 Randomized Controlled Trials for Medical Devices	3
1.4 Challenges of Modeling and Simulation for Medical Device Evaluation . . .	6
1.5 Chapter Overview and Contributions	7
CHAPTER 2 : Formal Specification of Cardiac Physiology and ICD Algorithms . .	15
2.1 Chapter Overview	15
2.2 Introduction	15
2.3 Electrophysiology of the Heart	16
2.4 Tachyarrhythmia and ICDs	18
2.5 Towards Formal Verification of ICDs	23
2.6 Chapter Conclusion	34
CHAPTER 3 : Problem Formulation and Related Work	36
3.1 Chapter Overview	36
3.2 Problem Formulation	36
3.3 Clinical Trial of a Hypothetical Device and Target Population	40
3.4 Historical Controlled Trials	46

3.5	In-silico Clinical Trials for Medical Device Evaluation	51
3.6	Main Approach and Key Questions	57
3.7	Chapter Conclusion	59
CHAPTER 4 : Virtual Cohort Generation and In-silico Medical Device Evaluation		60
4.1	Overview	60
4.2	Introduction	60
4.3	Virtual Cohort Generation	61
4.4	In-silico Evaluation of the ICD	68
4.5	Bayesian Hierarchical Models for Virtual Cohort Generation	74
4.6	Chapter Conclusion	80
CHAPTER 5 : Computer-aided Clinical Trials and Uncertainty Quantification . .		82
5.1	Overview	82
5.2	Introduction	82
5.3	Modeling Uncertainty and Assumptions in a CACT	84
5.4	Combining Virtual Cohorts and Real Patient Data	91
5.5	The δ -Robustness of Computer-Aided Clinical Trial (CACT) Outcomes . .	94
5.6	Case Study: CACT for Rhythm ID Goes Head-to-head Trial (RIGHT) (CACT- RIGHT)	98
5.7	Discussion	107
5.8	Chapter Conclusion	110
CHAPTER 6 : A Data-driven Approach for Improving ICD Performance		112
6.1	Overview	112
6.2	Introduction	112
6.3	Problem Formulation	114
6.4	Approach	115
6.5	Dataset Description	117
6.6	Evaluation	118

6.7	Results and Discussion	119
6.8	Chapter Conclusion	120
CHAPTER 7 : Conclusion		122
7.1	Summary of Contributions	122
7.2	Open Questions	123
7.3	Ending Remarks	125
APPENDIX		126
BIBLIOGRAPHY		129

LIST OF TABLES

TABLE 1 :	Examples of Physiological Models for In-silico Evaluation of Medical Devices	53
TABLE 2 :	Specificity and sensitivity of ICD VT/SVT discrimination algorithms	73
TABLE 3 :	Summary of RIGHT(Gold, Ahmad, et al., 2012) results.	104
TABLE 4 :	Default detection parameters for the Boston Scientific ICD Pipeline	116
TABLE 5 :	Class distribution for the EGM data set	117
TABLE 6 :	Default and optimized parameter settings for the detection rate thresholds	119

LIST OF ILLUSTRATIONS

FIGURE 1 :	Examples of life-critical, medical cyber-physical systems	2
FIGURE 2 :	Structure of a Clinical Trial for Medical Devices	3
FIGURE 3 :	Chapter Overview and Contributions	8
FIGURE 4 :	Electrophysiology of the heart	17
FIGURE 5 :	Phases of an action potential (AP)	17
FIGURE 6 :	VT and SVT cardiac signals	19
FIGURE 7 :	ICD operation	20
FIGURE 8 :	ICD hardware components	21
FIGURE 9 :	ICD algorithm pipeline	22
FIGURE 10 :	Cardiac tissue is modeled as a 2D grid of cells.	25
FIGURE 11 :	Hybrid model of one hybrid cellular automaton.	26
FIGURE 12 :	Sample outputs from CA model	28
FIGURE 13 :	Examples of CA Model Simulation	29
FIGURE 14 :	Hybrid automata of ICD sensing algorithm	32
FIGURE 15 :	RCT of a medical device and statistical power	37
FIGURE 16 :	Example of target population	42
FIGURE 17 :	Example Clinical Trial	43
FIGURE 18 :	Sample size vs. distribution of estimated $\hat{\theta}$	45
FIGURE 19 :	Comparison of sample size and power	46
FIGURE 20 :	Historical data variability vs. power	50
FIGURE 21 :	Benefits of in-silico evaluation.	52
FIGURE 22 :	Effects of simulation uncertainty on the outcome of the trial . . .	56
FIGURE 23 :	Comparison of methods by sources of information.	58
FIGURE 24 :	Simulation trace for timed-automata model	63

FIGURE 25 : TA model of electrical conduction system of the heart	64
FIGURE 26 : EGM morphology extraction example	66
FIGURE 27 : EGM generation process	67
FIGURE 28 : Hardware interface for in-silico evaluation of ICD	68
FIGURE 29 : Overview of an in-silico pre-clinical trial (ISPCT)	69
FIGURE 30 : Estimated rates of inappropriate therapy	72
FIGURE 31 : Effect of Duration and VF threshold parameters on specificity . .	75
FIGURE 32 : Distribution of Ventricular Cycle Length (VCL) for generated vir- tual cohort	78
FIGURE 33 : Results of Bayesian hierarchical model for virtual cohort generation	79
FIGURE 34 : Overview of a computer-aided clinical trial (CACT) for medical devices and robustness evaluation.	84
FIGURE 35 : Monitoring of CACT prior using a discount function.	92
FIGURE 36 : δ -robustness of outcome $H(\phi; \pi(\cdot), \alpha)$	96
FIGURE 37 : Example of generated physiological signal	99
FIGURE 38 : Example of marginalization using Monte Carlo methods.	101
FIGURE 39 : Pre-clinical simulation robustness evaluation	102
FIGURE 40 : Pre-clinical simulation robustness plane.	103
FIGURE 41 : Change in mean difference distribution	105
FIGURE 42 : Results of post-trial simulation and robustness evaluation for CACT- RIGHT.	106
FIGURE 43 : Post-trial analysis robustness plane.	107
FIGURE 44 : Comparison of CACT in terms of power vs. sample size	108
FIGURE 45 : Comparison of CACT in terms of power vs true θ	109
FIGURE 46 : Results of parameter search for the Weibull discount function. . .	110
FIGURE 47 : Rhythm ID detection pipeline	113
FIGURE 48 : Subpopulation Exclusion	114

FIGURE 49 : Data-driven device improvement with Bayesian Optimization . . .	115
FIGURE 50 : Default and Bayesian optimization parameter setting performance	118
FIGURE 51 : Parametric recall for the slow-VT sub-population	120

CHAPTER 1 : Introduction

1.1. Clinical Trials for Medical Treatment Evaluation

Life-saving medical treatments and interventions have advanced rapidly in the last century thereby increasing life-spans and improving overall quality of life. One driver of such accelerated medical innovation is the advance of clinical research methods, in particular the randomized controlled trial (RCT). In an RCT, a type of clinical trial, patients are randomly assigned to receive a specific treatment in a prospective study and the outcomes are compared against a control of standard treatment. In terms of classification by levels of evidence (Burns, Rohrich, and Chung, 2011), the results from an RCT is considered one of the most definitive types of evidence of safety and efficacy for an intervention compared to any other type of clinical research information. The randomization in an RCT reduces biases in outcomes due to confounding factors and heterogeneity in the patient cohort, leading to a highly accurate assessment of safety and efficacy.

This level of rigor comes at a cost as RCTs are massive endeavors which occur over prolonged periods of time and require careful planning to be as efficient. Despite the advances in tools and methods, protocol design remains challenging and clinical trials often fail to provide sufficient evidence of effectiveness (Fogel, 2018). A recent study by (Wong, Siah, and Lo, 2019) found that only 13.8% of drug development programs eventually lead to market approval. For high-risk devices requiring clinical data, from 2008 to 2017, only 1.5% obtained approval (Dubin et al., 2021). Moreover, even after approval, unexpected outcomes occur when the treatment is deployed to a more general, heterogeneous population. During the same period, of the approved high-risk devices 27.1% were recalled (Dubin et al., 2021).

In recent decades, advances in physiological modeling and the development of large-scale simulation methods has driven the interest in leveraging in-silico evaluation as part of the regulatory approval process for medical treatments. This thesis focuses on medical devices and utilizing such computational techniques in order to improve clinical trials.



(a) Pacemakers and ICDs



(b) Automated insulin pumps



(c) Cochlear implants

Figure 1: Examples of life-critical, medical cyber-physical systems

1.2. Limitations of Clinical Trials for Evaluating Life-critical Medical Devices

Medical devices, as all other medical treatments, must demonstrate a high standard of safety and efficacy in an RCT in order to gain market approval and be deployed to the target population. Medical devices have advantages in terms of the development process and approval for safety when compared to pharmaceuticals. Development of a device occurs in faster iterations compared to pharmaceuticals and the core functionality can be modeled and predicted to some degree. This leads to differences in the evaluation process for the safety and efficacy of a device (Faris and Shuren, 2017). Even though the U.S. FDA requires clinical evidence for high-risk devices, it allows a pathway for device approval that is considered largely equivalent to a previously approved device with minimal modifications (FDA, 2019). However, these advantages can also be an obstacle when trying to incorporate more innovative technologies as there can be significant discordance between the speed of device development and the time for approval when evaluated with clinical trials.

In the case of life-critical medical devices, such as the implantable cardioverter defibrillator (ICD), the limitations of current methods of clinical trials are exacerbated. In particular, ICDs are cyber-physical systems which monitor a patient’s state and actively intervene to prevent dangerous conditions. Fig. 1 provides examples of implantable devices which are cyber-physical systems. The software component enables the device to treat a heterogeneous population via configuration and tuning of settings. However, their configurability leads to added challenges when determining guidelines for using these devices which is based on evidence from clinical trials (Garnreiter, 2017).

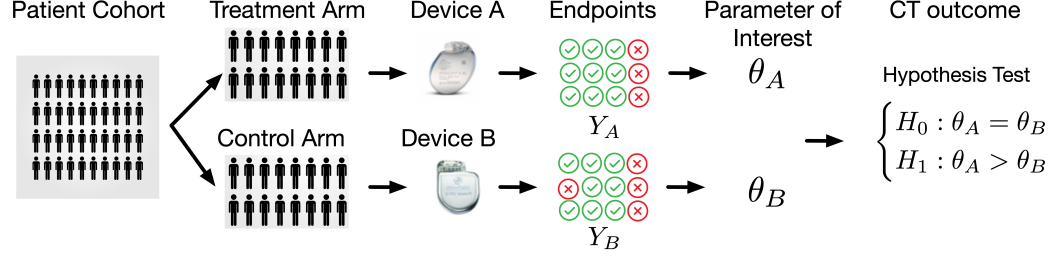


Figure 2: Structure of a clinical trial for medical devices. In a typical clinical trial, a patient cohort is randomly assigned to a treatment arm and a control arm. Each arm of the trial receives a different device or treatment. Based off the endpoints which are collected from arm, the clinical trial answers a hypothesis test with respect to a parameter of interest regarding the device performance.

As medical devices increase in complexity and incorporate more advanced technologies, such as components with machine-learning based algorithms, evaluating the safety and efficacy of such devices will be even more challenging. The state of these devices can potentially evolve over time which can improve the specificity of a device for an individual or subpopulation while increasing the risk of negative outcomes for a different subpopulation. Simply increasing the number and scope of trials would be costly and ineffective in establishing the safety of such devices. Methods to improve the prediction of clinical trial outcomes before they are executed and analyzing existing outcomes to allow for more pointed trials with greater efficiency are needed. With respect to this need, this thesis explores the following questions:

1. How can we use data generated from computer modeling and simulation in the context of a clinical trial and improve the evaluation of medical devices?
2. How can we account for the uncertainty in outcomes when using simulated data?
3. How can we improve device performance with scalable, data-driven methods?

1.3. Randomized Controlled Trials for Medical Devices

In order to further understand the limitations and the possible approaches for improving clinical trials, we first need to understand the basic components of an RCT. Fig. 2 depicts the structure of a typical clinical trial for medical devices. In a randomized controlled trial, an intervention or treatment for a disease is evaluated in comparison to a control. In the case of pharmaceuticals, this control is commonly a placebo whereas the treatment

is some new drug. Generally, a patient *cohort* is defined and recruited according to strict requirements necessary for evaluating an intervention. In the most basic form, a patient will be randomly allocated to either a *treatment* arm, where the patient will receive the new intervention, or a *control* arm, which receives either a placebo or another treatment according to some standard of care.

In medical device trials, for both practical and ethical reasons, control arms are rarely placebo (Faris and Shuren, 2017). Instead, the trial arms consist of patients receiving different devices or the same device with different software configurations. In order to satisfy an inquiry while maintaining a reasonable trial size, clinical trials are limited in the types of questions that can be asked and the various configurations of devices which can be tested. This leads to unexpected outcomes when deploying such software-based devices.

For each patient in a clinical trial, a measurable data point or *endpoint* is defined and collected over the course of a trial. For example, in a pharmaceutical clinical trial for a hypertension reduction treatment, the amount of reduction in blood pressure for patient can be measured before and after receiving the drug or a placebo. In the case of the ICD, the number of inappropriate therapies caused by the device during a trial can be measured as an endpoint. The endpoint is determined according to the *parameter of interest* such that some *sufficient statistic* can be computed to estimate the parameter of interest. Based off the estimate of the parameter of interest, a *hypothesis test* is used to determine the equivalence or superiority/inferiority of the treatment intervention relative to the control.

1.3.1. A clinical trial for a generic medical device

The goal of this thesis is to provide a method to generate an additional source of endpoints using computer modeling and simulation and incorporate it to improve the evaluation of a medical device. We illustrate these concepts through an example of a clinical trial for a generic medical device as shown in Fig. 2. For example, if the parameter of interest θ is the parameter of a Bernoulli distribution which governs a random variable $Y \sim \text{Bern}(\theta)$

where,

$$Y_i = \begin{cases} 1 & \text{positive treatment outcome for } i\text{th patient} \\ 0 & \text{negative treatment outcome for } i\text{th patient} \end{cases}$$

Here, θ would be defined as the *rate of positive outcomes*. Y would be whether or not a patient in a cohort has obtained an positive outcome. For a patient cohort of size N , the set of measurements of $\{Y_i\}$ comprise the set of endpoints. From these endpoints, it can easily be shown that the estimate of parameter, $\hat{\theta}$, can be computed as the expectation of Y ,

$$\hat{\theta} = E[Y] = \frac{\sum_{i=1}^N Y_i}{N},$$

where Y_i is the value of endpoint for the i th patient.

Let us define the rate of positive outcomes for the device in the treatment arm and control arm to be θ_T and θ_C , respectively. A clinical trial for evaluating the superiority of a treatment would want to determine if $\theta_T > \theta_C$. To derive this conclusion, a clinical trial could be designed as a hypothesis test of significance α with null hypothesis H_0 and alternative hypothesis H_1 defined as follows:

$$H_0 : \theta_T = \theta_C$$

$$H_1 : \theta_T > \theta_C + \epsilon$$

Here, $\epsilon > 0$ is the *effect size* or difference in the parameter of interest. The outcome of the hypothesis test would be to either reject the null hypothesis in favor of the alternative hypothesis and conclude that $\theta_T < \theta_C$ with significance α . Here, the significance level α and the effect size ϵ are all factors that are considered to determine the necessary cohort size or sample size N of a clinical trial.

The significance level α is the chance of rejecting the null when in fact the null is true (i.e. a false positive) and is an important factor in determining the quality of a clinical trial.

In addition to the the significance level, the power $Power(1 - \beta) = 1 - \beta$, where β is the FNR (false negative rate), is an important factor in determining design of a clinical trial and the quality of the outcome. The power of a hypothesis test determines the chance that a significant result will be detected and the null hypothesis will be rejected assuming that the alternative hypothesis is true. Therefore, having a higher power is desirable.

The quality of a clinical trial outcome can be determined by the significance α and the power. Both of these factors are effected by the cohort sample size N and the effect size ϵ . Intuitively, it is easier to detect a significant difference between the treatment and control arm when ϵ is larger and would require a smaller sample size N for the same α and β . In comparison, if ϵ is small, then a larger sample size N will be needed for the same α and β . This trade-off between the necessary sample size and the desired significance α and power β is one of the key factors in determining a clinical trial design as well as one of the deciding factors in the success of a clinical trial.

1.4. Challenges of Modeling and Simulation for Medical Device Evaluation

Conventional approaches for improving clinical trial design and execution include alternative trials designs, such as adaptive designs(Mahajan and Gupta, 2010), and safety mechanisms, such as planning interim analyses with the possibility of early-stopping (Evans, 2010). However, given the increasing complexity and development speed of devices, adhering to conventional methods will be limited in improving the evaluation of devices. As mentioned previously, increasing the sample size or conducting additional clinical trials would not only be infeasible economically but unethical and ineffective.

This thesis aims to utilize computer modeling and simulation of physiology and devices to generate *virtual cohorts* which can be used as an alternative source of information to predict outcomes of a clinical trial for medical devices and improve the statistical power during evaluation and improve the performance of the device. With respect to these objectives, the following challenges are addressed:

1. First, this thesis addresses the challenge of how to generate a virtual cohort to use in the context of a clinical trial.
2. The next challenge is then how to systematically incorporate the virtual cohort in order to improve the evaluation of a medical device in a clinical trial.
3. Third, since simulation is bound to have inaccuracies, how to account for the uncertainty when analyzing results from a clinical trial evaluation is addressed.
4. Finally, the potential for an automated, data-driven method for improving the performance of device itself is explored.

1.5. Chapter Overview and Contributions

Fig. 3 depicts the main contributions of this thesis and the related chapters. In what follows, we give an overview of the chapters and the corresponding contributions:

Chapter 2: Formal Specification of Cardiac Physiology and ICD Algorithms

Chapter 2 introduces critical background on cardiac physiology and pathology. While there are many causes of cardiac disease, the effects of cardiac disorders often manifest as arrhythmia which impair the body's ability to supply oxygen to cells resulting in morbidity. Importantly, not all observable changes in heart rate or rhythm are dangerous; for example, an athlete may experience increased heart rate during a competition. Thus, we describe the taxonomy of dangerous conditions which require medical intervention. Then, given this set of clinically relevant conditions, we introduce a treatment used to resolve them. In particular, we highlight the ICD, a class of implantable device, which detects and resolves dangerous tachyarrhythmia via the application of short, high-energy shocks. As with any detection based system, the performance is predicated on the ability of the device

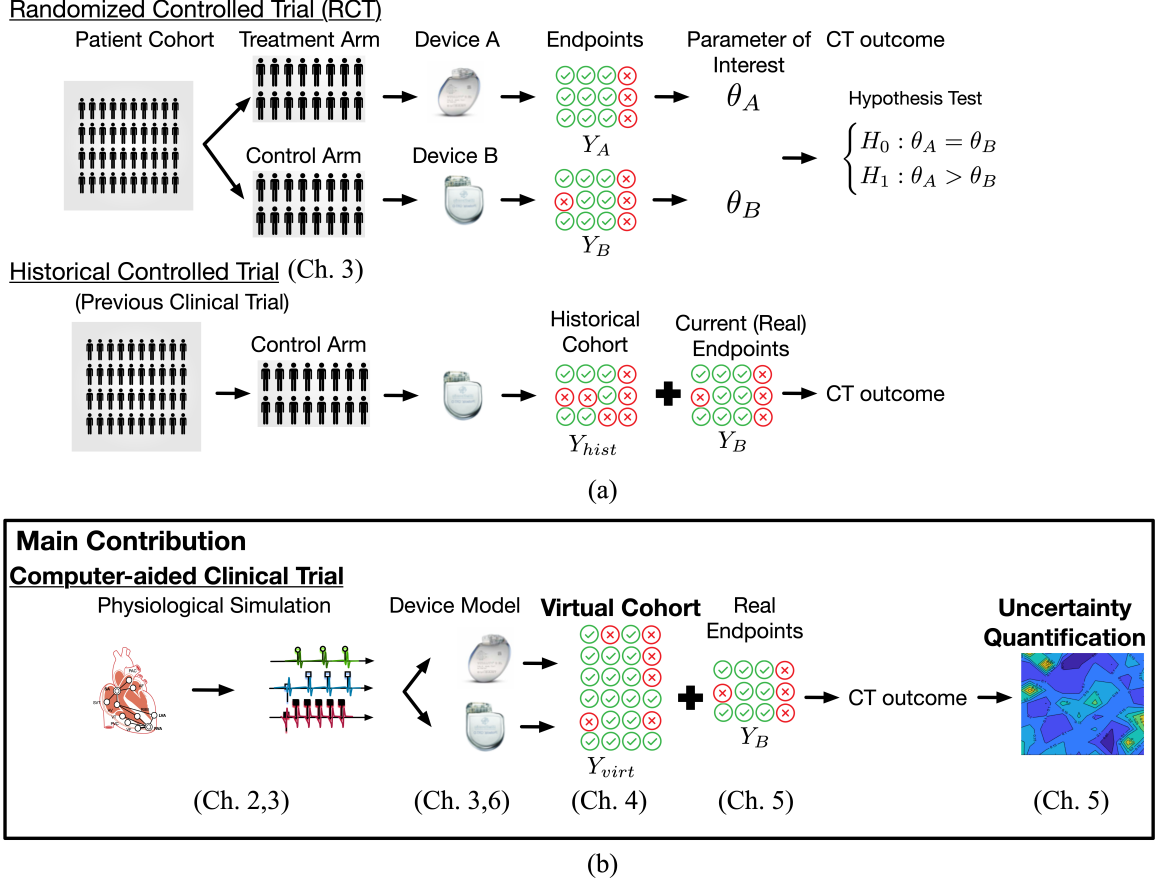


Figure 3: Chapter Overview and Contributions. Computer-aided clinical trials are presented in comparison to two alternative settings: (a) historical controlled trials (b) physiological simulators for device evaluation. Chapters relevant to each contribution are highlighted.

to correctly resolve true positive incidences of dangerous tachyarrhythmias while correctly ignoring other nominal, or non life-threatening conditions.

Unlike pharmaceuticals, implantable medical devices such as ICDs contain software components which implement, in an algorithmic fashion, a mapping from measurements of the state of the heart to a treatment decision. The correctness of such devices can be evaluated on several levels. First, we may be interested whether an abstract model of the combined system can ever reach a state in which treatment is not applied when the condition of interest is present (similarly, if treatment can be applied when the condition is not present). Second, as the design of the device progresses and abstract details are concretized it is

necessary to evaluate how the implementation performs when subject to *in-situ* conditions (H. Abbas, Mittelmann, and Fainekos, 2014). Due to the limitations of the device, *e.g.* power consumption, real-time constraints, signal noise, etc., the device may differ from the abstract initial design, requiring further evidence of its suitability for its intended use.

While the majority of this thesis focuses on the second question, estimating the performance of the device implementation, we motivate its importance by first exploring the limitations of verifying performance with *abstract* models of closed-loop ICD systems. Due to the non-linear, hybrid dynamics of the ICD’s interaction with cardiac physiology, it is first necessary to investigate whether formally checking the properties of ICD performance is even *decidable* (Henzinger et al., 1998). The main contribution of this chapter is, surprisingly, a positive answer to the question of decideability. Nevertheless, it is via this positive result that the computational, modeling, and framing challenges of the ICD, and, more broadly, medical device evaluation, can be cast in a sharper light.

Despite the feasibility of verification, an important difference exists between verifiable properties of the system and the performance of the device in the context of a clinical trial. The guarantees provided by verification are with respect to the abstracted model and require conformance of a particular patient’s physiology. The clinical reality is that identifying such parameters for a cohort of patients or even a single patient is difficult. Although verification techniques could be used early in the design phase of an ICD algorithm, employing alternative models which capture the distribution of patient physiology and incorporating statistical hypothesis testing methodologies is needed. Thus, for the rest of this thesis, we endeavor to utilize the actual device code to make assertions about the safety within the context of a clinical trial and develop an appropriate simulator for generating samples of the necessary physiological signals.

Related publications:

1. Houssam Abbas, Kuk Jin Jang, Zhihao Jiang, and Rahul Mangharam (Apr. 2016).

- “Towards Model Checking of Implantable Cardioverter Defibrillators”. In: *Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*. HSCC ’16. New York, NY, USA: Association for Computing Machinery, pp. 87–92. ISBN: 978-1-4503-3955-1. DOI: 10.1145/2883817.2883841
2. Houssam Abbas, Kuk Jin Jang, and Rahul Mangharam (Feb. 2017). “Nonlinear Hybrid Automata Model of Excitable Cardiac Tissue”. en-US. in: *EPiC Series in Computing*. Vol. 43. ISSN: 2398-7340. EasyChair, pp. 1–8. DOI: 10.29007/5zfk
 3. Samuel Huang, Madeline Diep, Kuk Jin Jang, Elizabeth M. Cherry, Flavio H. Fenton, Rance Cleaveland, Mikael Lindvall, Rahul Mangharam, and Adam Porter (Nov. 2020). “Towards Automated Comprehension and Alignment of Cardiac Models at the System Invariant Level”. In: *CSBio ’20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*. CSBio2020. New York, NY, USA: Association for Computing Machinery, pp. 18–28. ISBN: 978-1-4503-8823-8. DOI: 10.1145/3429210.3429225

Chapter 3: Problem Formulation and Related Work

In Chapter 3, we begin by formally defining RCTs. Using a simple example, which includes a hypothetical blackbox medical device and parameterized distribution describing a target population, we illustrate the implications of treatment effect, variability of the target population, and sample size on the power of RCTs. We highlight the relation between sample size and the power of a clinical trial; unsurprisingly, larger sample sizes have more power. However, simply recruiting more patients to increase the sample size would be unethical as the safety and efficacy of a device is unproven in a prospective RCT.

Thus, we investigate two alternative settings: historical controlled trials and physiological simulators. As depicted in Fig. 3(b), historical controlled trials approach the challenge of limited sample size by assuming that data from control arms in prior trials can be utilized in a prospective trial, thereby allowing more allocation of the cohort to the treatment arm.

This can lead to biases under-powering the clinical trial in cases where a large variability exists in the population.

One of the key challenges of using a historical control is that, by definition, such methods only augment the control arm of a the experiment and cannot address the differences in patient composition, diversity in outcomes, and size of the treatment arm. In contrast, a patient generated from a physiological simulator may be assigned to the control or the treatment arm preserving the randomization property of the RCT. Thus, we show through the use of an ideal simulator that RCTs for medical devices could be designed to have arbitrarily high power. However, as George Box once pointed out (Box, 1976), the existence of a perfect model or simulator is *de-facto* impossible. Therefore, with this motivation, this thesis seeks new methods which enable the use of imperfect simulation data in the context of an RCT. The aim is to be able to incorporate simulation data and combine it with real data observed in a clinical trial, while providing metrics which help establish the credibility of the outcomes despite the inaccuracy of the simulator.

Chapter 4: Virtual Cohort Generation and In-silico Medical Device Evaluation

Chapter 4 develops a method for generating a virtual cohort of synthetic signals using a simulator. The main contribution of this chapter is the development of a generative model for physiological simulation of cardiac signals called *electrograms* (EGM). The model enables the creation of virtual cohorts which can be used in concert with ICD discrimination algorithm implementation artifacts to simulate trial endpoints.

The quality of the proposed model is investigated empirically. In our experiments we demonstrate the ability of the model to predict results aligned with an actual clinical trial, the Rhythm ID Goes Head-to-Head Trial (RIGHT, Gold, Ahmad, et al., 2012). In RIGHT, contrary to the expected result, the treatment devices demonstrated a higher risk of inappropriate therapy compared to the result. Predicting this outcome and re-evaluating the feasibility of the trial would have lead to immense savings in trials costs and efforts. Moreover, we demonstrate prediction of secondary outcomes that agree with findings in sub-

sequent clinical trials. Nevertheless, the initial analysis demonstrated that while the results obtained from the proposed model are directionally accurate they still included some bias regarding the characteristics of the physiological signal, and thus failed to converge to the true mean performance observed in the trial. In order to improve the prediction, we further refine the proposed model by re-defining it as a hierarchical model that allows the incorporation of historical information which would reduce the bias in the signal characteristics. This improved the prediction outcomes for RIGHT by approximately 42%.

Through this empirical investigation, we were able to discern some of the shortcomings of in-silico evaluation, namely, that assumptions made during the virtual cohort generation process can greatly affect the overall outcome. This is an important factor when considering the credibility of the evidence provided with in-silico evaluation. The next chapter presents the main approach for addressing this limitation.

Related publications:

4. Houssam Abbas, Zhihao Jiang, Kuk Jin Jang, Marco Beccani, Jackson Liang, and Rahul Mangharam (Oct. 2016). “High-level modeling for computer-aided clinical trials of medical devices”. In: *Proc. IEEE Int. High Level Design Validation and Test Workshop (HLDVT)*, pp. 85–92. DOI: 10.1109/HLDVT.2016.7748260
5. Zhihao Jiang, Houssam Abbas, Kuk Jin Jang, Marco Beccani, Jackson Liang, Sanjay Dixit, and Rahul Mangharam (2016). “In-silico pre-clinical trials for implantable cardioverter defibrillators”. In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the. IEEE*, pp. 169–172

Chapter 5: Computer-aided Clinical Trials and Uncertainty Quantification

A major challenge of using in-silico evaluation is how to combine real and virtual cohorts while accounting for the uncertainty in the results due simulation inaccuracies. Quantifying this uncertainty would allow for more informed decision-making when using virtual cohorts in the evaluation of a medical device. To address these issues, for the first contribution of

this chapter, we develop a statistical framework which we call the computer-aided clinical trial (CACT), which allows us to incorporate the virtual cohort as an additional source of information and allows for explicit modeling of assumptions and sources of uncertainty.

For the next contribution, the framework is applied in the case study of RIGHT where a significant reduction in sample size and improvement in power is demonstrated when compared to historical controlled trials. As the final contribution of this chapter, we define δ -robustness as a measure of quantifying the uncertainty in the outcomes of a clinical trial due to simulation assumptions and propose a method for approximating the value.

Related publications:

7. Kuk Jin Jang, James Weimer, Houssam Abbas, Zhihao Jiang, Jackson Liang, Sanjay Dixit, and Rahul Mangharam (July 2018). “Computer Aided Clinical Trials for Implantable Cardiac Devices”. eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2018*, pp. 1–4. ISSN: 2694-0604. DOI: 10.1109/EMBC.2018.8513284
8. Kuk Jin Jang, Yash Vardhan Pant, Bo Zhang, James Weimer, and Rahul Mangharam (Apr. 2019). “Robustness evaluation of computer-aided clinical trials for medical devices”. In: *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems. ICCPS '19*. New York, NY, USA: Association for Computing Machinery, pp. 163–173. ISBN: 978-1-4503-6285-6. DOI: 10.1145/3302509.3311058

Chapter 6: A Data-driven Approach for Improving ICD Performance

In the preceding chapters, this thesis examines how to better measure the performance of implantable medical devices like ICDs through the use of simulation. While these techniques can increase the power of the hypothesis tests used to determine the safety and efficacy of the device, they do not provide a direct method to improve the performance of the device if it is found to be lacking. Thus, in Chapter 6, we address the challenge of improving the

performance of the ICDs using a data-driven approach. In particular, we focus on how the discrimination algorithm in the device can be improved through the identification of better device configurations. Today, device settings are tuned in an ad-hoc fashion by clinicians.

The primary contribution of this chapter is the demonstration of an approach for obtaining improved ICD algorithm parameters using Bayesian optimization. Empirically, our experiments show improved accuracy, precision, and recall of the ICD detection algorithm investigated in previous chapters by 5.%, 4.4%, and 6.5%. The results segue into a deeper discussion about optimization objectives and the concept of algorithmic fairness which we discuss in greater depth in the concluding chapter.

Chapter 7: Conclusion and Open Problems in Computer-aided Clinical Trials

In this final chapter, we summarize the overall contributions of this thesis. In addition, a discussion about open problems in computer-aided clinical trials is presented, including those regarding the regulatory aspects of in-silico evaluation and algorithmic fairness. The contributions of this thesis provides a statistical framework and methods that allow for the improvement of the development and evaluation process of systems in medicine and beyond.

In what follows, we present our solution for generating virtual cohort and utilizing it as a source of information to predict clinical trial outcomes and improve evaluation as well as our approach for improving device performance. In the next chapter, we begin by introducing details with respect to cardiac physiology and the main domain application of the ICD.

CHAPTER 2 : Formal Specification of Cardiac Physiology and ICD Algorithms

2.1. Chapter Overview

In this chapter, we provide background on cardiac physiology and critical details regarding medical device evaluation and the main application domain, the implantable cardioverter defibrillator (ICD). Elements of this chapter have been adapted from “Towards Model Checking of Implantable Cardioverter Defibrillators” in the Proceedings of International Conference on Hybrid Systems: Computation and Control 2016, “Nonlinear Hybrid Automata Model of Excitable Cardiac Tissue” in the Workshop for Applied Verification for Continuous and Hybrid Systems 2016, and “Towards Automated Comprehension and Alignment of Cardiac Models at the System Invariant Level” in the Proceedings of the Conference on Computational Systems-Biology and Bioinformatics 2020. These papers were joint work with Houssam Abbas, Zhihao Jiang, Samuel Huang, Madeline Diep, Elizabeth Cherry, Flavio Fenton, Rance Cleaveland, Mikael Lindvall, Adam Porter, and Rahul Mangharam.

2.2. Introduction

2.2.1. Safety evaluation of medical devices.

The process of evaluating medical devices differs from pharmaceuticals. Device development generally occurs at a faster pace than pharmaceuticals with iterations/improvements occurring throughout the lifecycle of the device.

When compared to pharmaceuticals, the physics of the device are better understood and the operation can be more easily modeled. In the case of cyber-physical systems, such as the ICD, the computational, software component plays a major role in the operation. The software component implements an algorithm which maps measurements of the state of the heart to a treatment decision. The operating characteristics of these devices can be readily simulated and allows for correctness and safety characteristics to be evaluated at several levels during the various stages of the development process. At a first level, one can evaluate an abstract model of the combined system to determine if the system can

reach some undesired state. This type of evaluation would typically be applied at early stages of the design in order to verify the correctness and safety of the algorithm. Second, as the device design progresses, it becomes necessary to evaluate how the implementation of the system performs under in-situ conditions with the actual physiological signals or within a the real environment of the patient. This is because the implementation may be significantly different from the initial abstract design due to physical constraints such as power consumption, real-time constraints, language expressiveness, sensor quality, etc. At these latter stages, the evaluation of the device would be typically be measured in the context of a clinical trial.

2.2.2. Chapter Summary

The methods and approaches developed in this thesis focus on the second level of evaluating medical device in a clinical trial and present the results within the context of a evaluating the ICD. Although most of this thesis focuses on the evaluation of the device within the context of the clinical trial, we motivate the importance by first exploring the limitations of verifying an abstract model of the ICD and cardiac physiology. In what follows, we first demonstrate that verification is possible and conclude with a discussion about the abstractions developed and the resulting limitations of using a verification-based approach.

2.3. Electrophysiology of the Heart

The human heart maintains blood circulation through a concerted contraction of the atria and ventricles triggered by an electrical signal controlled by the cardiac conduction system depicted in Fig. 4. In a normal heart beat, an electrical signal begins in a specialized region of the heart, called the sinoatrial (SA) node, and spreads through the atria. Subsequently the signal conducts through the atrio-ventricular (AV) node, His-Purkinje network, and finally propagates through the ventricles (Fogoros, 2017). The electrical signal that passes through the heart is produced by cardiac events at the cellular level known as *action potentials*.

The action potential of an excitable cardiac cell results from the exchange of various ionic currents across the cell membrane at varying magnitudes in response to the change in

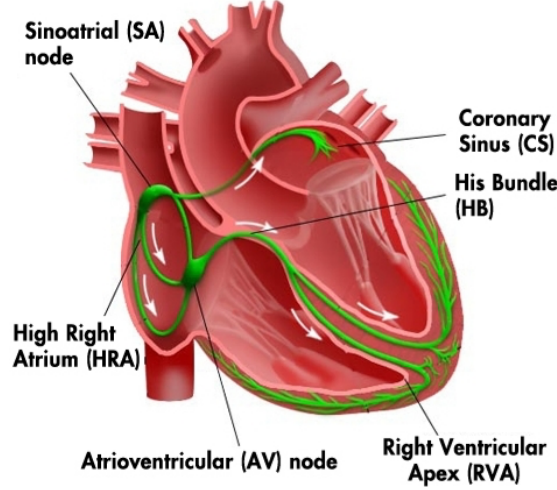


Figure 4: Electrophysiology of the heart. During a contraction of the heart, an electrical signal starts from the SA node and spreads through the atria, AV node, his-purkinje network, and then through the ventricles (Image from (Huang et al., 2020))

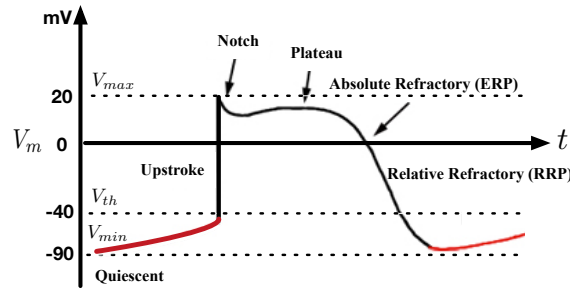


Figure 5: Phases of an action potential (AP). AP figure from (Hood, 2015)

voltages, resulting in a net depolarization or repolarization. Typically, a cell is at a baseline voltage. When the potential exceeds a set threshold, various gating mechanisms and ionic currents are activated in sequence, the net result of which is the measured action potential. Fig. 5 depicts the phases of an action potential. The action potential can be divided into various phases beginning with an upstroke of rapid depolarization (phase 0) during which the muscle contracts. Following depolarization is an early repolarization phase (phase 1), followed by the plateau (phase 2), which allows the muscle to hold its contraction. This plateau is characteristic of cardiac cells and differs from other excitable cells, such as nervous cells, which also exhibit action potentials. After the plateau, repolarization occurs at a higher rate resulting in the downstroke (phase 3), when the muscle relaxes and refills. This

phase can be divided into the absolute refractory phase (ERP) where additional stimuli cannot cause another depolarization and the relative refractory phase (RRP) during which an additional depolarization can occur given an appropriate stimulus. Finally, the cell reaches the resting phase (phase 4), where the transmembrane voltage of the cell reaches a steady state. An action potential is triggered by an increase in voltage above some threshold caused by an action potential propagating from neighboring tissue or from an artificial pacing signal, as in the case of cardiac devices such as a pacemaker.

In most clinical settings and in medical devices, action potentials are not directly measured and are instead observed as electrograms (EGM) or electrocardiograms (ECGs), which we can be simulated as pseudo-ECGs. Pseudo-ECGs can be computed as $ECG = \int (\frac{D\nabla V \cdot \vec{r}}{r^3}) dr$, where r is the vector from the recording electrode to a point of measurement (Clayton et al., 2011). Intuitively, the electrogram is a weighted sum of the surrounding potentials and represents a higher-level observation of the underlying activity of the heart. The main difference between ECGs and EGMs is that ECGs are measured on the surface of the patient and represent a more global picture of the entire heart. In comparison, EGMs are measured through intra-cardiac leads implanted within the chambers of the heart. There are two main types of EGM, one of which is the bi-polar EGMs and the other is the uni-polar EGMs. Uni-polar EGMs are measured from some reference point (e.g. the can of the device) to the electrode at the measurement point. Bi-polar EGMs measure the potential difference between two electrodes and is equivalent to the potential difference of two uni-polar EGMs.

2.4. Tachyarrhythmia and ICDs

2.4.1. Types of tachyarrhythmia

EGMs and ECGs are used to diagnose cardiac arrhythmia which are irregular heart rhythms caused by abnormalities of the heart's electrical conduction system. ICD therapy targets a specific class of arrhythmia called tachyarrhythmia or tachycardia which are rhythms at an elevated heart rate above the normal resting rate or normal sinus rhythm (NSR). Broadly, there are two main types of tachycardia which are of interest in this thesis. These

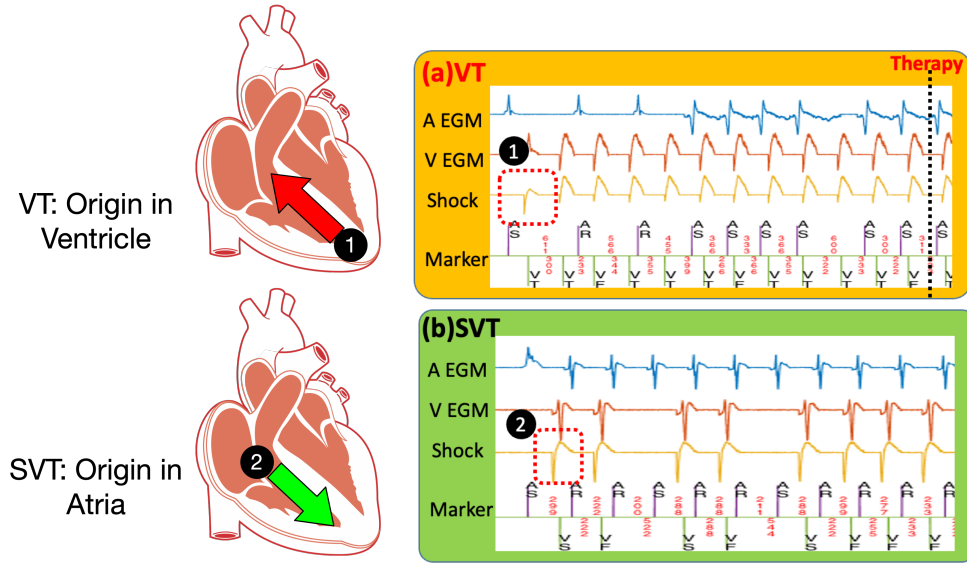
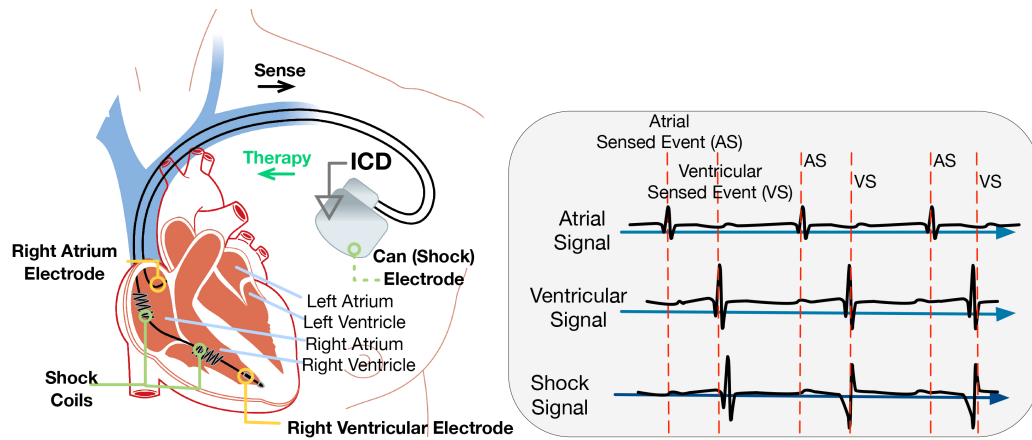


Figure 6: Visualization of VT and SVT cardiac signals according to their location of origin in the heart. (a) VT originates in the ventricles and is characterized by a fast ventricular heart rate. (b) SVT originates in the atria and is characterized by a fast atrial heart rate. Notice the morphology differences between the VT and SVT signals highlighted by the red box in the shock channel, where (1) indicates an abnormal signal morphology often representative of a VT episode. (Image credit: Zhihao Jiang)

rhythms are categorized by the origin of the irregular electrical activity within the heart: supraventricular tachycardia (SVT) and ventricular tachyarrhythmia. Examples of the cardiac signals for each of these tachycardia are included in Figure 6.

SVT occur above the ventricular, hence the name 'supra-ventricular', and begins in the atria or along the atrial conduction pathways. These arrhythmia occur when either regions in or near the atria trigger an abnormal activation or maintain an irregular conduction, thus interfering with the regular impulse and conduction initiated by the SA node. The abnormal atrial contractions speed up the atrial heart rate and can lead to an elevated ventricular rate. In the case of SVT, episodes tend to be unpleasant rather than life-threatening and treatment with ICD therapy may not be necessary. However, the issue is that these rhythms may be mistaken by device algorithms for ventricular tachyarrhythmia.

In comparison, ventricular tachyarrhythmia originate in the ventricles and is characterized by a rapid ventricular heart rate. Specific types of ventricular tachyarrhythmia include



Rate of Inappropriate Therapy

PRL+W: 54.1%
Rhythm ID: 62.2%

Figure 7: ICD operation. The ICD consists of the generator (can) and leads which are implanted into the chambers of the heart. The ICD monitors the signal called electrograms (EGMs) and detects depolarization events. These events are interpreted through the algorithm of the ICD. In the RIGHT trial, the Rhythm ID algorithm in Vitality 2 devices was compared to Medtronic devices with the PR-Logic + Wavelet (PRL+W) algorithm with respect to the risk of inappropriate therapy. Inappropriate therapy rates reached as high as 62.2% and 54.1% for each of the device algorithms, respectively.

ventricular tachycardia (VT) and ventricular fibrillation (VF). VT occurs when the beating of the ventricles is no longer controlled by the SA node but rather by another electrical impulse along the lower electrical pathway within the ventricles. VT is a fast rhythm, but still maintains some regularity in the shape or morphology of the waveform at a faster rate. However, VF has a much faster ventricular rate than VT as electrical impulses are initiated in multiple locations within the ventricle leading to a disconcerted contraction of the ventricle. The resulting chaotic heartbeat pumps blood inefficiently and requires immediate therapy. In either case, VTs and VFs are life-threatening as they prevent the proper blood circulation throughout the body, leading to sudden cardiac arrest and death. Immediate medical attention and termination of such rhythms with treatments such as ICD defibrillation are critical for the survival of the patient.

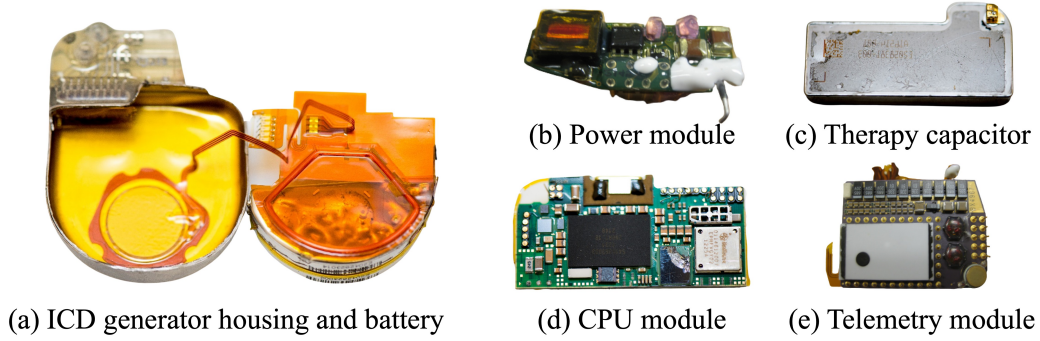


Figure 8: ICD hardware components. (a) ICD generator(can) housing which encases the remaining components, the connectors for the leads, and the battery. An RF antenna allows for remote configuration and transfer of data during follow-ups. (b) Power module (c) Capacitor which is used to create the charge when therapy is necessary. (d) The main CPU module which implements the OS and main discrimination algorithm. (e) Telemetry module for RF communication. (Image credit: Marco Beccani)

2.4.2. Overview of ICD system components

The implantable cardioverter defibrillator (ICD) is a life-saving medical device which detects tachycardia and intervenes by automatically applying a high-energy electrical shock when required. This device is small enough to be implanted under the skin and autonomously monitors a patient’s heart rhythm. For non-ventricular tachycardias, such as SVT, treatment may not be necessary, whereas VTs require immediate therapy (Moss et al., 2012).

The ICD consists of the generator, which contains the battery and houses the computational components which includes the core discrimination algorithm, and multiple wires called leads, which are implanted in the heart. The leads are connected to the generator on one end which acts as a reference point and the cardiac tissue on the other as shown in Fig. 7. Through these leads, EGMs are measured to monitor the electrophysiological activity of the heart as explained previously. Fig. 8 shows the hardware breakdown of an ICD. The generator housing and battery take up most of the volume of the ICD (Fig. 8(a)). An RF antenna can be seen in the housing and works with the telemetry module (Fig. 8(e)) for remote configuration and data communication during check-ups. The power module (Fig. 8(b)) provides power for ICD operation and the capacitor (Fig. 8(c)) is charged in order to apply a defibrillation shock. The core CPU module (Fig. 8(d)) controls ICD operation and

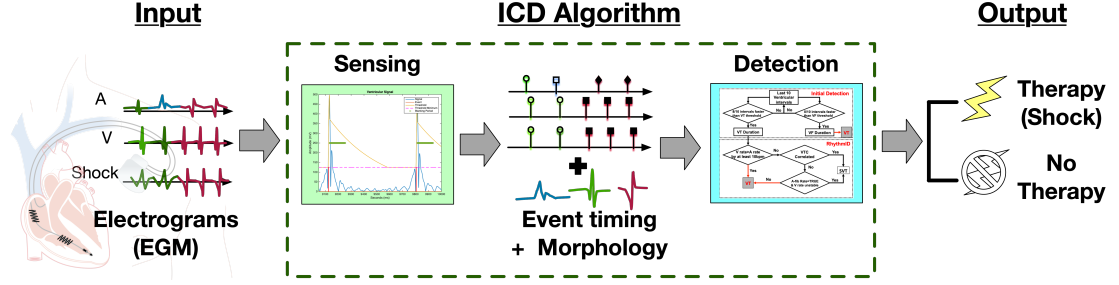


Figure 9: The ICD algorithm consists of two stages, sensing and detection, and receives an electrogram signal recorded by the ICD as input. The sensing stage transforms the signal into a series of timing events and waveform morphology. This is used by the detection stage to detect abnormal rhythms. When an abnormal rhythm is detected, therapy is applied if it is identified as VT or inhibited if it is identified as an SVT or non-VT.

implements the discrimination algorithm.

As shown Fig. 9, the input EGM signal is monitored by the ICD algorithm and when an abnormal heart rhythm is detected, the generator sends electrical pulses to the heart through the leads in an attempt to terminate the irregular rhythm. These applied electrical pulses can be either anti-tachycardia pacing, a series of small rapid pulses that override the current rhythm, or a high-energy electrical shock, called defibrillation, to terminate the rhythm by restarting all conduction within the heart. In modern devices, the ICD also incorporates other functionality to target other classes of rhythms, such as a pacemaker which stimulates the heart when the rhythm is too slow. We cover the operation of the ICD in greater detail in Sec. 2.5.3.

2.4.3. Limitations of ICD therapy

Despite the proven life-saving capabilities of the ICD to prevent sudden cardiac death, the device has some limitations. At times it delivers inappropriate therapy. Inappropriate device-delivered therapy activations are harmful therapy activations delivered during a non-VT event and are typically caused by the misclassification of non-VT events as VT events. As shown in the Rhythm ID Goes Head-to-head Trial (Gold, Ahmad, et al., 2012), inappropriate therapy rates reached as high as 62.2% and 54.1% for each of the respective devices in the trial. These inappropriate therapies can increase the risk of mortality in

patients (Moss et al., 2012; Rees et al., 2011). Therefore, it is essential to evaluate the safety of the device in terms of the rate of inappropriate therapy.

2.5. Towards Formal Verification of ICDs

Evaluating the correctness and safety of the ICD algorithm is essential in order for ICD therapy to be effective. The nature of the device allows for various levels of evaluation, depending on the stage of development. During the early stages of design, it may be of interest to determine if the ICD algorithm will ever reach an undesired state, such as when it applies inappropriate therapy. This type of question can potentially be answered through formal verification, but the feasibility must first be determined. Therefore, in the remaining sections we explore the necessary steps for verifying properties of abstract models of closed-loop ICD performance and determine the feasibility of such an approach. We first develop a cardiac model of the tissue using hybrid systems approach for modeling cardiac tissue. Next, we present an overview of how the ICD can be modeled as a hybrid system and in particular a STORMED hybrid system.

2.5.1. Preliminaries

The non-linear dynamics of cardiac tissue can be modeled as a hybrid system. Some preliminaries regarding hybrid systems can be found in Appendix A.1. In general, hybrid systems are a robust formalism that can model many classes of dynamical systems. Unfortunately, except for a relatively specific subset of problems, hybrid systems are undecidable and verification is infeasible (Alur et al., 2000). STORMED hybrid systems are a special class of hybrid systems for which verification is possible. The definition and characteristics of STORMED hybrid systems can be found in Appendix A.2 and (Vladimerou et al., 2008). The defining characteristics of STORMED hybrid systems is the following theorem:

Theorem 1. (Vladimerou et al., 2008) Let \mathcal{H} be a STORMED hybrid system, and let \mathcal{P} be an o-minimal partition of its hybrid state space. Then \mathcal{H} admits a finite bisimulation that respects \mathcal{P} .

Theorem 1 implies that if a system can be shown to be a STORMED hybrid system, then the verification of formally-defined properties of the system may be feasible. Therefore, in this section, we first develop a hybrid systems model of cardiac tissue and later demonstrate how the cardiac model of the system is a STORMED hybrid system. For this, we now present some background necessary for modeling cardiac activity as a hybrid system.

2.5.2. Cardiac modeling and hybrid automata model of cardiac tissue

Overview of cardiac modeling

A large spectrum of heart models have been developed differing in complexity, fidelity, represented species and region of the heart, and properties that the models can express (F. H. Fenton and Cherry, 2008). On one end, there are many models that describe cardiac action potentials by directly tracking detailed ion currents across the membrane and changes to ionic species (Ten Tusscher and Panfilov, 2006; O’Hara et al., 2011) and a number with simplified ionic currents (F. Fenton and Karma, 1998; Mitchell and Schaeffer, 2003). On the other end, some models further abstract the cellular or tissue behavior using finite-state automata such as cellular automata (Saxberg and Cohen, 1991; Makowiec, Wdowczyk, and Struzik, 2019), timed automata (Z. Jiang, Pajic, and R. Mangharam, 2012), or hybrid automata (Yip et al., 2016).

Hybrid automata model of cardiac tissue

The model of cardiac tissue that we discuss in this section is based on *cellular automata (CA)*, which we formalize as nonlinear hybrid automata. Cellular automata have been widely used for modeling biological systems (Ermentrout and Edelstein-Keshet, 1993). In (Spector et al., 2011), authors demonstrated its ability to simulate meaningful cardiac phenomena such as ectopics (irregular isolated beats) and re-entrant tachycardias (which is a common class of potentially fatal tachycardias). It has also been used to study the measurement process of ICDs. The model we develop in this section follows the description in (Spector et al., 2011) and modifies it slightly to make the resulting waveform more realistic, as described in the following sections.

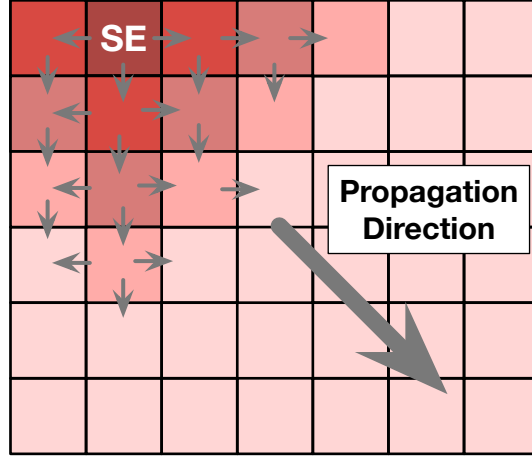


Figure 10: Cardiac tissue is modeled as a 2D grid of cells. *SE* is a self-exciting cell. After *SE* depolarizes, the neighboring cells depolarize as well. The delay in propagation is determined by the velocity of depolarization, how long the cell remains depolarized, the resistance to current flow between cells, and the current state of the neighboring cell. The chain reaction of depolarization causes an aggregate wave of AP propagation.

Cellular Automata Model

See Fig. 5. Initially, the cell is in a *quiescent*, polarized state where the membrane potential is at a *resting potential*. The typical resting potential is about $V_m = -90mV$. The complex interactions within a neighborhood of the cell allow the possibility that a net influx of current can occur within the cell causing V_m to rise. If V_m rises above a threshold value V_{th} , an AP is triggered. $V_{th} = -40mV$ in a typical cardiac myocyte. The cell enters a *depolarization* phase where the cell's voltage V_m rapidly increases. V_m increases until a maximum potential V_{max} is reached (nominally around $56mV$), at which point the cell begins an *initial repolarization* phase. This phase can be represented as a 'notch' in the signal. Due to the ion-channel interactions at the cellular level, cardiac myocytes demonstrate an extended, slower repolarization phase called a *plateau* phase. Afterwards, a phase of rapid repolarization occurs. The repolarization phase can be further divided into an *absolute refractory* phase, where the cell is unreactive to external stimuli, and a *relative refractory* phase, where external stimuli can cause an additional AP of lesser magnitude. Finally the cell returns to its initial fully repolarized state.

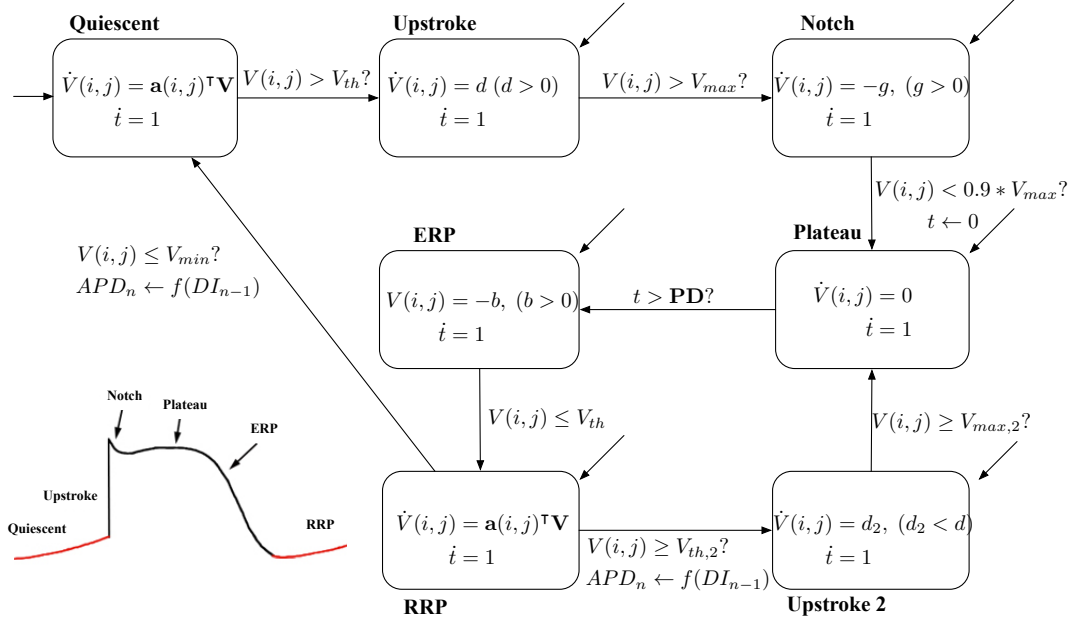


Figure 11: Hybrid model of one hybrid cellular automaton. AP figure from (Hood, 2015). $V_{th,2} > V_{th}$, $V_{max,2} < V_{max}$. DI_n is the Diastolic Interval of n^{th} beat, and f is the restitution function which determines the action potential duration (APD) for the subsequent cycle based on the diastolic interval (DI) of the current cycle.

The excitable heart tissue is composed of individual cells arranged in a 2D grid of $N \times N$ as shown in Fig. 10. Cells interact with each other via a four-neighborhood structure. Each cell is modeled as a nonlinear hybrid automaton. The continuous state of cell (i, j) is $[V_m(i, j), t_{ij}]$, where $V_m(i, j)$ is the cross-membrane voltage and t_{ij} is a local timer. The cell automaton has 7 modes, which model the 7 phases of an AP. See Fig. 11. They are *Quiescent*, *Upstroke*, *Notch*, *Plateau*, *absolute refractory period (ERP)*, *relative refractory period (RRP)*, and *Secondary upstroke*. We now describe the dynamics in all modes.

Quiescent. Initially a cell is in the quiescent mode. Typically, $V_m(0) = V_{min} = -90mV$ in this mode. The $(i, j)^{th}$ cell's voltage at time t in this phase depends on that of its 4

neighbors and its own as follows (Spector et al., 2011):

$$\begin{aligned}
\dot{V}_m(i, j, t) &= V_{intr} + \frac{[V(i-1, j, t) + V(i+1, j, t) - 2V_m(i, j, t)]}{R_h(i, j)} \\
&\quad + \frac{[V_m(i, j-1, t) + V_m(i, j+1, t) - 2V_m(i, j, t)]}{R_v(i, j)} \\
&= V_{intr} + a(i, j)^\top \mathbf{V}(t), \quad a(i, j) \in \mathbb{R}^{N^2}
\end{aligned} \tag{2.1}$$

where R_h, R_v are resistance constants that can vary across the myocardium. In Quiescent mode, $V_{intr} = 0$ for most excitable cells (ECs) whereas $V_{intr} > 0$ for a self-exciting cell. $\mathbf{V} = (V(1, 1), \dots, V(N, N)) \in \mathbb{R}^{N^2}$ contains all voltages in the grid.

Upstroke - Depolarization. In Upstroke, the voltage increases exponentially according to $\dot{V}(i, j) = d > 0$.

Notch - Initial Repolarization. Upon reaching V_{max} , the voltage decreases slightly per $\dot{V}(i, j) = -g < 0$.

Plateau. While the cell is in the plateau mode, V_m remains constant for a given duration PD (Plateau Duration). Biologically, the delayed reaction time of slower Ca^{++} ion channels causes the plateau. In a more realistic AP, the plateau is not constant but decreases slightly.

ERP - Absolute Refractory Period. Next, the cell begins a secondary repolarization phase which can be divided into two phases, the first of which is ERP. During this mode, the cell is resilient to external stimuli which is reflected in update equations for the state.

From Upstroke to the end of ERP, the cell can not be excited by its neighbors. This is reflected in the dynamics, which depend solely on the intrinsic voltage of the cell.

RRP - Relative Refractory Period After ERP, the cell enters the last phase of repolarization, the RRP mode. During this period, the cell is susceptible to current flows from neighboring cells. In this mode, the dynamics follow Eq. (2.1). If the voltage increases above a threshold $V_{th,2} > V_{th}$ due to the interactions with its neighbors, the cell can enter a

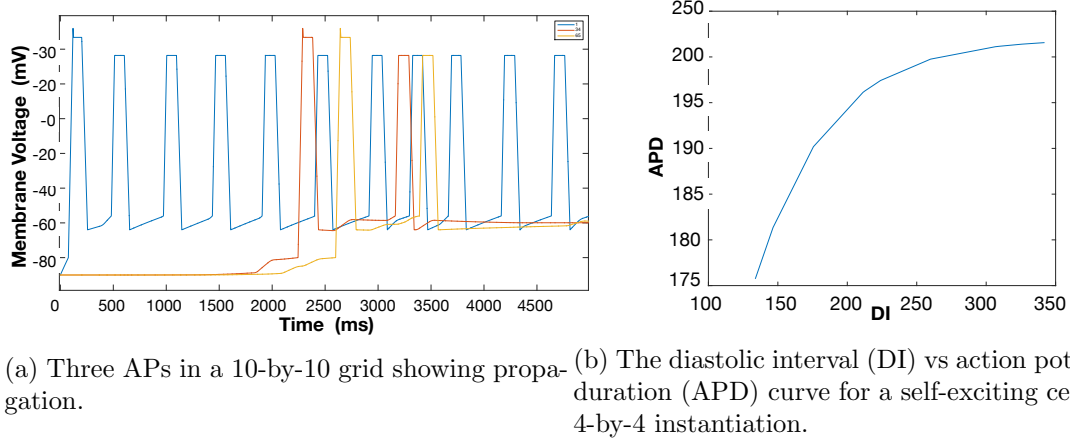


Figure 12: Sample outputs from CA model

secondary depolarization mode, Secondary upstroke. If this occurs, the cell depolarizes to a voltage $V_{max}' < V_{max}$ albeit with a smaller slope. If on the other hand, the voltage goes back to the quiescent level, the cell enters Quiescent.

Simulation outputs

In this section, we present some example outputs of the implementation of the model. Fig. 12a shows 3 APs from 3 non-contiguous cells in a 10-by-10 grid: a self-exciting cell (at position (1,1)) and two excitable cells from the middle of the grid, at positions (4,4) and (5,7). As can be seen the AP travels from the self-exciting cell (which is the first to depolarize) to its neighbors.

The restitution curve is an important feature of cardiac tissue, and is responsible for the non-linearity of this model. Broadly speaking, it gives the duration of the next AP, known as APD, as a function of the Diastolic Interval DI_{n-1} which lasts from the end of the previous AP and the current upstroke. We measured the successive (DI, APD) pairs for cell (1,1) and plotted the resulting curve. Fig. 12b shows that the simulated curve matches the shape of the experimentally obtained curves in vivo. Finally in Fig. 13 we show the progression of the electrical wave in an inhomogeneous tissue (i.e., whose resistance changes spatially) from three self-exciting cells in the lower left corner.

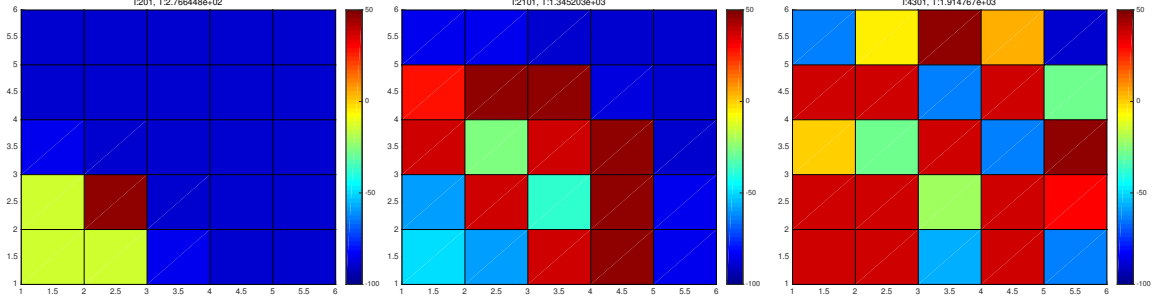


Figure 13: Three time snapshots of tissue, showing a depolarization wave propagating, left to right. Warmer colors indicate a more recent upstroke. Because the tissue is inhomogeneous, propagation does not proceed uniformly across the tissue, whence the observed eventual fractionation (last panel on the right).

Observations and limitations of the model

The first key observation is that the large number of parameters in this model ($18n^2$ for an n -by- n grid) makes it very challenging to select values that lead to desired phenomena. E.g., simply sustaining a propagating wavefront is not trivial: if we choose upstroke slopes too large, then the AP duration decreases progressively which can compromise propagation. If the upstroke velocities are too small on the other hand, cell voltages may never exceed the depolarization threshold a second time and the tissue is electrically dead. This highlights the need for parameter synthesis in this model and others like it (Bogomolov et al., 2015). We also emphasize that obtaining desired phenomena is also a matter of neighborhood structure, and depend on the restitution curve.

Another observation is that in this model the transitions of the various automata can be extremely close in time, since cells that are electrically near will naturally synchronize with each other. This can create numerical issues for ODE solvers. E.g. with Matlab’s ode45 (which implements Runge-Kutta (4,5) method), in mostly homogeneous tissue, a few mode switches were either doubly detected leading to fake transitions, or incorrectly reported as being duplicate and thus transitions were missed. We have written code to detect some of these cases, but we feel that such an issue is best dealt with by the solvers themselves, e.g. by the usage of verified integrators. Mode switches are also very frequent as cells go

through their APs slowing simulation. E.g., on a 2.2 GHz, 16 GB Intel Core i7, simulating 6 seconds of a 6-by-6 grid took an average 872secs, and an 8-by-8 grid took 2252secs.

Despite these limitations, the model demonstrates the potential for modeling cardiac tissue as a hybrid system to be used for the formal verification of ICDs, which we discuss next.

CA model of cardiac tissue is a STORMED Hybrid system

A key result regarding this model is that the model can be shown to be a STORMED hybrid system and thus admit a finite bisimulation. The whole heart model \mathcal{H}_{CA} is the parallel composition of these N^2 single-cell models. The $(i, j)^{th}$ cell's voltage at time t in Phase 4 depends on that of its neighbors and its own as follows

$$\begin{aligned}\dot{V}(i, j, t) &= \frac{[V(i-1, j, t) + V(i+1, j, t) - 2V(i, j, t)]}{R_h(i, j)} \\ &\quad + \frac{[V(i, j-1, t) + V(i, j+1, t) - 2V(i, j, t)]}{R_v(i, j)} \\ &= a(i, j)^T V(t), \quad a(i, j) \in \mathbb{R}^{N^2}\end{aligned}\tag{2.2}$$

where R_h, R_v are conduction constants that can vary across the myocardium. Thus V evolves according to a linear ODE $\dot{V} = AV$ where A is the matrix whose rows are the $a(i, j)$. The two states t and t_p are clocks. Clock t_p keeps track of the value of the last discrete jump. We will use this arrangement in all our models: it avoids resetting the clocks which preserves Reset Monotonicity.

As mentioned previously, implantable cardioverter defibrillators (ICDs) observe through channels of signals called an electrogram (EGM) signal. The signal read on a channel can be modeled by Correa de Sa et al., 2011:

$$s(t) = \frac{1}{K} \sum_{i,j} \left(\frac{1}{\|p_{i,j} - p_0\|} - \frac{1}{\|p_{i,j} - p_1\|} \right) \dot{V}(i, j, t)\tag{2.3}$$

where $\|\cdot\|$ is the Euclidian norm, p_0 and p_1 are the electrodes' positions and $p_{i,j}$ is the position of the $(i, j)^{th}$ cell on the 2D myocardium ($p_0, p_1, p_{i,j} \in \mathbb{R}^2$). Positions p_0, p_1 should

be chosen different from $p_{i,j}$ to avoid infinities.

From this definition of the model and the signals observed we can conclude the following:

Theorem 2. Let \mathcal{H}_{CA} be the whole heart cellular automaton model obtained by parallel composition of N^2 models \mathcal{H}_c with state vector $x = [V, t, t_p, s] \in \mathbb{R}^{N^2} \times \mathbb{R}^3$. Assume that all executions of the system have a duration of $D \geq 0$. Then \mathcal{H}_{CA} is STORMED.

The proof follows from the definition of a STORMED hybrid system and the reader is directed to (H. Abbas, Jang, and Rahul Mangharam, 2017). Since the publication of this model, several other groups have presented models using similar approaches. A in-depth review is beyond the scope of this thesis.

Next, we will continue to modeling the ICD algorithm as a hybrid system.

2.5.3. Formal specification of ICD discrimination algorithms

Following from the results in the previous sections, we now demonstrate how the ICD algorithm itself can be modeled as a hybrid system.

The cardiac model can be used to generate EGM signals which would then be observed by the ICD. The ICD processes the signal in two stages where first the measured signal is converted into timing events (Fig. 7). As shown in Fig. 9, the timing events along with the morphology of the signal observed are used in the next stage, called *discrimination*. During this stage, various components, called *discriminators*, are defined and employed to determine the state of the patient being observed. In (H. Abbas, Jang, and Rahul Mangharam, 2017), the various components of the sensing and discrimination algorithm modeled as a hybrid system and shown to each be a STORMED hybrid system. The different components of the algorithm shown as hybrid systems can be found in the (H. Abbas, Jang, and Rahul Mangharam, 2017).

Sensing algorithm as a hybrid system

As an example, we describe the hybrid system model of the sensing algorithm in an ICD.

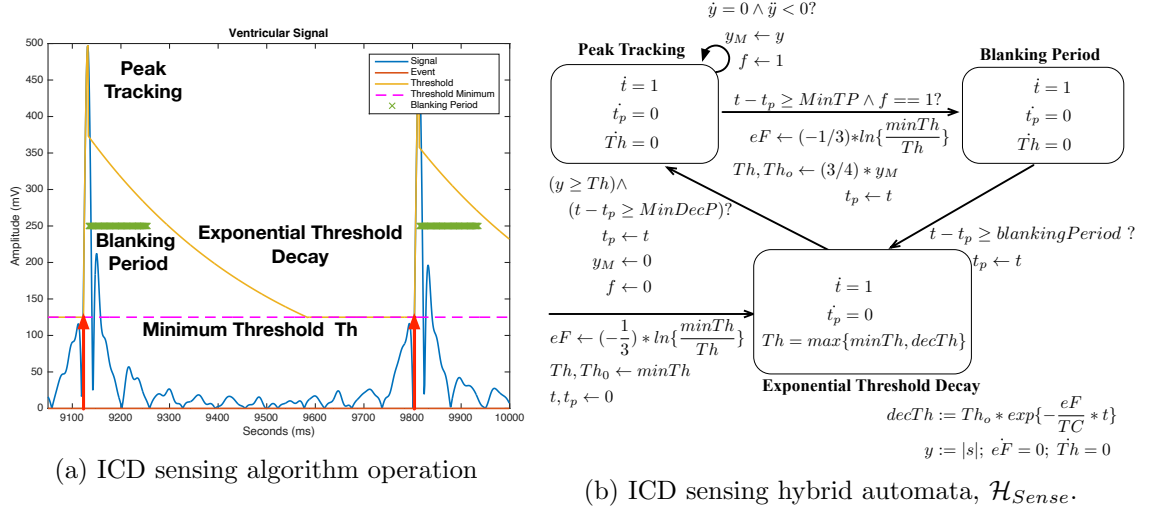


Figure 14: (a) Example of dynamic threshold adjustment in ICD sensing algorithm. The shown signal is rectified. (b) \mathcal{H}_{Sense} . States not shown in a mode have a 0 derivative, e.g., $e\dot{F} = 0$ in all modes.

Sensing is the process by which cardiac signals s measured through the leads of the ICD are converted to timing events. The ICD declares events when the signal exceeds a dynamically-adjusted threshold Th . Fig. 14a illustrates the operation of an ICD sensing algorithm. After an event has been declared, the algorithm begins to track the peak of the event. At the same time, the a blanking period is initiated to prevent the double-counting of the same event. Afterwards, the threshold is dynamically adjusted until it reaches the minimum threshold. The threshold is dynamically adjusted in order to operate robustly in complex environments where cardiac events can vary greatly in signal amplitude and frequency, such as during VF.

Fig. 14b shows the corresponding hybrid-automata model \mathcal{H}_{Sense} of the sensing algorithm. The sensing takes place on the rectified EGM signal $y = |s|$. After an event is declared at the current threshold value ($y(t) \geq Th(t)$ in Fig. 14b), the algorithm tracks the signal in order to measure the next peak's amplitude (Peak Tracking). For a duration $MinTP$ (min tracking period) the latest peak is saved in y_M . A variable f indicates that a peak was found. After a peak is found ($f == 1$) and after the end of the tracking period, the algorithm enters a fixed *Blanking Period* (Blanking), during which additional events are

ignored. On the transition to Blanking, Th , Th_0 and the exponential factor of decay eF are updated. At the end of the blanking period, the algorithm transitions to the Exponential Decay mode in which Th decays exponentially from Th_0 to a minimum level (Exponential Decay), and stays there for at least a sampling period of $MinDecP$. Different manufacturers may use a step-wise decay instead of exponential, but the principle is the same. Local peak detection is modeled via the $\dot{y} = 0 \wedge \ddot{y} < 0$ transition. While $y = |s|$ is non-differentiable at 0, the peak will occur away from 0, as shown in Fig. 14a. States t, t_p are clocks and $minTh$ and TC are constant parameters. In (H. Abbas, Jang, and Rahul Mangharam, 2017), it is shown that each of the various components of the discrimination algorithm can be modeled as a STORMED hybrid system.

2.5.4. Composition of ICD model with cardiac model

The composition of each of these components of an ICD algorithm can be shown to be a STORMED hybrid system. The details can be found in (H. Abbas, Jang, and Rahul Mangharam, 2017), but the main result is replicated here:

Theorem 3. Let $\Sigma_i = (\mathcal{H}_i, \mathcal{A}, \phi^i, b^{i,-}, b^{i,+}, d_{min}^i, \varepsilon^i, \zeta^i)$, $i = 1, \dots, m$ be deterministic SHS defined using the same underlying o-minimal structure, and where each state space X^i is bounded by B_{X^i} .

Define parallel composition $\Sigma = (\mathcal{H}, \mathcal{A}, \phi, b^-, b^+, d_{min}, \varepsilon, \zeta)$ where $\mathcal{H} = \mathcal{H}_1 || \dots || \mathcal{H}_m$, $\phi = (\phi^1, \dots, \phi^m)^T \in \mathbb{R}^{mn}$, $b^{i,-} = \inf_{x \in X} \phi \cdot x$, $b^{i,+} = \sup_{x \in X} \phi \cdot x$, $\varepsilon = \min(\min_i \varepsilon^i, \min_i \frac{\zeta^i}{B_{X^i}})$, $\zeta = \min_i \zeta^i$ and $d_{min} = \min_{I \subseteq [m]} (\min_{i \in I} d_{min}^i, \min_{i \in I, j \in [m] \setminus I} d_{min}^{ij})$. Assume that the following

Collection Separability condition holds: for all $i, j \leq m, i \neq j$ there exists $d_{min}^{ij} > 0$ s.t. if $x \in X$ is in the reachable set of \mathcal{H} and $x^i \in G_e^i \wedge x^j \notin G_{e'}^j \forall e' \in E^j$ then $d(x^j, G_{e'}^j) > d_{min}^{ij}$ for all $e' \in E^j$ where E^j is the edge set of Σ_j and $G_{e'}^j$ is a guard of Σ_j on edge $e' \in E^j$. Then Σ is STORMED.

From here, it can be shown that if an approximate reachability tool with definable over-approximations is available for the continuous dynamics, it can be used in (A.1) to yield a

finite *simulation* for the STORMED hybrid system.

In summary, Theorem 3 and the results of the previous section allow us to conclude that the CA model of cardiac tissue and hybrid system ICD model is a STORMED hybrid system and thus, the composition is also a STORMED hybrid system. This positive result demonstrates the feasibility of formal verification of the closed-loop ICD performance.

2.5.5. Current limitations to formal verification of ICDs

Despite the feasibility of verification, the cardiac signals produced from this model cannot be directly applied to software artifacts implementing ICD algorithms. Hence, there exists an important gap between verifiable statements regarding the abstract model and the performance of the implementation in reality. This gap is exacerbated by the clinical difficulties associated in identifying the conformance of a particular patients physiology with the proposed model; simply, it is impossible to measure and identify the properties of each cell within a patients heart tissue. As mentioned in Sec. 2.5.2, the large number of parameters in this model ($18n^2$ for an n -by- n grid) makes it very challenging to select values that lead to desired phenomena and capture the behaviors of a patient.

While verification techniques could be used early in the design phase of new ICD algorithms, clinically relevant determinations of device safety and efficacy need to provide guarantees with respect to an entire target population and their behaviors. This will need to employ models which capture the distribution of patient physiology and incorporate statistical hypothesis testing methodologies. Moreover, the operating characteristics of the actual device will differ from the abstract model of the ICD. Thus, for the rest of this thesis, we endeavor to utilize the actual device code to make assertions about the safety within the context of a clinical trial.

2.6. Chapter Conclusion

In this chapter, we introduced the necessary concepts regarding the main application of the approaches in this thesis, mainly the electrophysiology of the heart and the ICD. As a

parallel branch of work, formal verification can serve as an alternative approach to evaluating the safety. In order to apply such a technique, it is necessary to have a model of the physical system and the device. We demonstrated how the cardiac activity in the heart and the operation of the ICD can be modeled as a STORMED hybrid system.

In the remainder of this thesis, we will describe the approach of utilizing simulated data as an additional source of information to aid in the evaluation of an ICD in the context of a clinical trial. The models presented here have the potential to be utilized within a clinical trial context. However, to do so would require additional efforts of data collection and evaluation in order to identify proper parameters and validate the generated signals. To better understand the requirements in a simulator for a clinical trial, in the following chapter, we first describe in more detail the process of a clinical trial and the challenges regarding clinical trial evaluation. This builds to the motivation behind the key approach of this thesis.

CHAPTER 3 : Problem Formulation and Related Work

3.1. Chapter Overview

In this chapter, we formalize the evaluation of the device in terms of a randomized controlled trial (RCT) for the scope of this thesis and define the objective. We introduce the related approaches to this thesis, in particular historical controlled trials and in-silico device evaluation. A comparison of the various approaches is demonstrated in the context of a hypothetical medical device. Finally, we describe the motivation and approach to resolving the central problem of this thesis of improving the power of a clinical trial through the use of physiological modeling and simulation to generate a virtual cohort.

3.2. Problem Formulation

3.2.1. Problem setup

Recall the structure of an RCT for a medical device from Fig. 2 replicated in Fig. 15. We can define a medical device as the function which takes as input the data X , such as some physiological signal, and outputs a decision Y . The decision is determined by X and some parameter(s) which manifest as a characteristic θ of the device, such as the inappropriate therapy rate of an ICD:

$$Y = f(x, \theta). \quad (3.1)$$

Here, X is a random variable from which is governed by the following distribution g :

$$X \sim p_X(x|z, \gamma), \quad (3.2)$$

where Z is an unobservable latent variable and γ is some observable variable. The fact that the variable is observable also means that it can be controlled or that various assumptions can be placed regarding the values of the variable. For a medical device, such as an ICD, an example of the physiological signal X would be the electrogram (EGM) which is input into

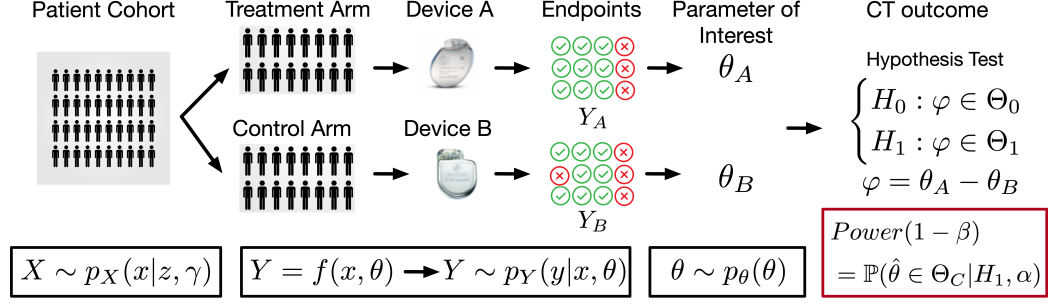


Figure 15: RCT of a medical device and statistical power. In this chapter, we formally define the RCT of Fig. 2. The first objective of this thesis is to improve the statistical power of a clinical trial for medical devices highlighted in red. See Sec. 3.2.

the device and Y would be the decision to apply a therapy or not. For instance, Y could be a diagnosing whether or not a patient is in a dangerous state and therapy needs to be applied by the device. In this example, we will assume that the appropriate behavior of the device can be determined. Even though the behavior of the medical device is deterministic, we can model Y as a random variable which is governed by some distribution conditioned on X and θ ,

$$Y \sim p_Y(y|x, \theta) \quad (3.3)$$

If we wish to account for the uncertainty in θ , we can define an optional prior on θ :

$$\theta \sim p_\theta(\theta), \quad (3.4)$$

where θ may again be conditioned upon the data X . The observations of Y in a clinical trial are called the *endpoints*.

3.2.2. Quality of a clinical trial

Our goal is to evaluate the performance of the device within the context of a clinical trial. Typically, in a clinical trial some quantity of interest, θ , will be selected regarding the performance of the device. For example, in a clinical trial comparing two devices A and B , θ_A can be the inappropriate therapy rate of one type of ICD while θ_B can be the inappropriate

therapy rate of the other ICD. For a clinical trial regarding a single continuous glucose monitor, θ can represent the change in average time a patient remains in hypoglycemia.

We define the RCT for a device as a *hypothesis test* with respect to θ . For example, when comparing two devices, we can define a null hypothesis, H_0 , where there is no difference between the two devices. We can represent this by defining a set $\Theta_0 = \{\theta_A | \theta_A = \theta_B\}$ as the set of values for H_0 . Equivalently, we can define $\varphi = \theta_A - \theta_B$ as the parameter of interest and $\Theta_0 = \{\varphi | \varphi = 0\}$. Similarly, we can define $\Theta_1 = \{\theta_A | \theta_A \geq \theta_B + \epsilon\}$ as the set of values for the alternative hypothesis, H_1 . In terms of φ , $\Theta_1 = \{\varphi | \varphi \geq \epsilon\}$. Here, ϵ is the effect size and is the difference in inappropriate therapy rates for the devices. We can define a hypothesis test of *significance level* α , such that,

$$H_0 : \varphi \in \Theta_0 \tag{3.5}$$

$$H_1 : \varphi \in \Theta_1 \tag{3.6}$$

where, the significance level α is synonymous with the false-positive rate of the trial or the Type I error rate. The hypothesis test determines if there is enough evidence to reject the null (i.e. $\theta_A = \theta_B$) in favor of the alternative hypothesis. In this example, rejecting the null means that the hypothesis test concludes that the inappropriate therapy rates of the two devices are not equivalent.

There are two representative quantities which are commonly used to evaluate the quality of a trial design and the outcome. First, there is the significance level of the test. As mentioned before, this is the false-positive rate of the trial and is defined as the probability of observing an extreme outcome in favor of the alternative hypothesis when the null hypothesis is true. This quantity is typically incorporated into the design of a trial and is defined according to the degree of rigor needed by the evaluation.

The second quantity is the *statistical power* the test. The statistical power of a hypothesis

test is defined as,

$$Power(1 - \beta) = \mathbb{P}(H_0 \text{ is rejected} | H_1 \text{ is true}) = 1 - \beta. \quad (3.7)$$

Here, β is the false negative rate or the Type II error rate. In words, it is the probability of rejecting the null hypothesis when the alternative hypothesis is actually true. Intuitively, it can be thought of as the rate at which the test will be able to detect a significant difference and conclude that the a null hypothesis false. While the significance level is often a fixed requirement which must be targeted, the power can be optimized further depending on the circumstances of the trial. Furthermore, it is possible for a trial to be determined to be over- or under-powered at the conclusion of the trial. Therefore, we focus on the statistical power of the trial as the quantity of interest that we would like to improve.

3.2.3. Overall objective: Improving the power of a clinical trial for medical devices

Whether or not H_0 is rejected is determined by the value of some *test statistic* $\hat{\theta} = T(Y)$ is in some set of critical values Θ_C . Here, T is a function computed from the observed endpoints Y which are governed by (3.3). This Θ_C depends on how H_0 and H_1 are defined for the hypothesis test. For example, when comparing the two devices, the estimated difference in the rates $\hat{\theta}$ can be a statistic and $\Theta_C = \{\hat{\theta} | \hat{\theta} > \delta\}$ can be defined as the critical region. In this case, one would reject the null hypothesis if $\hat{\theta}$ exceeds some threshold δ .

We can define the power of a clinical trial in terms of the test statistic and critical regions:

$$\mathbb{P}(\hat{\theta} \in \Theta_C | H_1, \alpha) \quad (3.8)$$

Here, H_1 means that the alternative hypothesis is true and α is the significance level.

The first goal of this thesis is to improve the power of a clinical trial for a medical device. For this, we must first understand the factors which can affect the power of a clinical trial.

For a given significance level α , the power tends to be greater when:

1. the effect size ϵ is large
2. the sample size N is large
3. the variances of the populations being sampled, $\sigma(Y)$, are small

Typically, in a clinical trial, ϵ and $\sigma(Y)$ are properties of the underlying data distribution or device and *cannot be controlled*. The sample size N is one of the most important variables in a clinical trial design which *can be controlled*.

For this reason, many approaches to improving RCTs have focused on the sample size in order to improve the power. Some approaches, for instance, efficiently determine and allocate the sample size in a trial design, while others work to improve the sample size by utilizing additional sources of information.

The following sections outline some approaches to improving upon the RCT by utilizing additional sources of data and achieve equivalent or greater power with a smaller sample size. A comparison of various approaches will be presented in the context of a hypothetical device and population.

3.3. Clinical Trial of a Hypothetical Device and Target Population

3.3.1. Overview

In order to illustrate the various challenges of improving statistical power, we further develop the generic device presented in the previous section as a toy problem of evaluating a hypothetical ICD device with a hypothetical population. We first introduce the device and then the target population of the device.

This example will illustrate the limitations of evaluating a device with a clinical trial related to sample size. One of the main limitations is that the sample size of a trial may not be sufficient relative to the effect size of a treatment. This can lead to negative results such as insufficient power of a trial or insignificant p-values. Another limitation is that certain subpopulations of the general population may be excluded from a clinical trial, resulting in

a biased result. This becomes problematic as not only are the results of clinical trials used to approve a device for deployment, the guidelines for treatment are based off the results. Biases in the guidelines can lead to negative outcomes for patients in subpopulations which may have been undersampled during clinical trials.

For the first limitation, historical controlled trials have been proposed as an alternative approach for evaluating medical treatments, in particular those of rare diseases. In a historical controlled trial, unlike a randomized controlled trial, the control arm is partially or completely replaced with outcomes from prior trials relative to the current trial. We demonstrate the advantages and limitations of a historical-controlled trial through the evaluation of the hypothetical medical device over the hypothetical target population.

3.3.2. Problem setup

In this problem we define two devices, $f_A(x, \theta_A)$ and $f_B(x, \theta_B)$, which we wish to evaluate in comparison to each other. These devices take as input the physiological signal modeled as a random variable, $X \sim p_X(x, \gamma)$, similar to before.

The endpoint is the output Y_D resulting from applying a sample of the physiological signal to each device, where $D \in \{A, B\}$ for each device. θ_D for each device is the parameter of interest regarding the performance of the device which we would like to compare. For example, θ_D can be the inappropriate therapy rate for device D .

Target population

We define a hypothetical target population as a one-dimensional mixture of Gaussians. The dimension corresponds to the ratio of average SVT ventricular rate to average VT ventricular rate x and ranges from 0 to 2. Fig. 16 depicts a sample distribution of the population that arises when sampling from such a distribution.

Generally speaking, the device will be able to correctly diagnose VT or SVT for a patient which is on the tail ends of the distribution and closer to 0 or 2. The device will incorrectly diagnose more episodes when the average rate of the patient is closer to 1. This is expected as

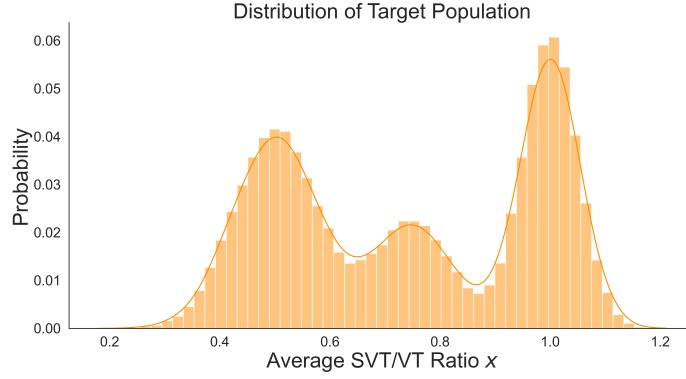


Figure 16: Target population. One-dimensional, mixture of Gaussian which correspond to the ratio of the avg. SVT ventricular rate to the avg. VT ventricular rate for patient. The figure depicts an example of the sample distribution from the target population. The device has a higher chance for inappropriate therapy when the ratio for a patient is between 0.5 and below 1.0. This implies that the ventricular rate of SVT is similar to ventricular rate of VT for that patient.

having similar ventricular rates makes it difficult for the device to discern the two conditions.

Device specification

For a device, f_D , the output is modeled as a Bernoulli random variable $Y_D \sim \text{Bern}(\theta_D)$ where,

$$Y_D = \begin{cases} 1 & \text{if inappropriate therapy occurred} \\ 0 & \text{o.w.} \end{cases} \quad (3.9)$$

In addition, Y_D is conditioned on whether x is greater or less than some threshold x_o ,

$$Y_D|x = \begin{cases} 0 & \text{if } x < x_o \\ \text{Bern}(\theta_D) & \text{if } x \geq x_o \end{cases} \quad (3.10)$$

In other words, if x is below the threshold x_o , the device will always be correct and above the threshold, and the probability of an inappropriate treatment is θ_D for the device.

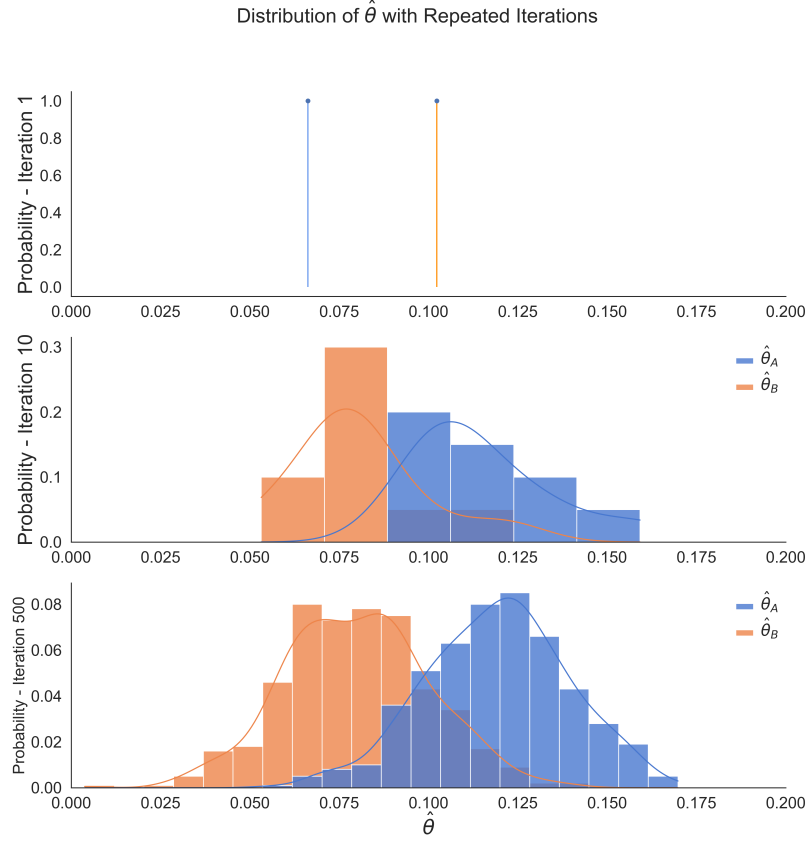


Figure 17: Example CT. (Top) A single trial results in a single estimate $\hat{\theta}_D$ for each device. (Middle) As trials are repeated, the estimates form a distribution around the true value of $\hat{\theta}_D$. (Bottom) After repeated trials, a distribution of the $\hat{\theta}_D$ can be observed. The difference in performance will be reflected in the degree of separation between the two distributions.

3.3.3. Clinical trial of a hypothetical medical device

In the clinical trial, we assume that the two devices have different values of θ_D . Additionally, we will assume the device A has a higher inappropriate therapy rate than device B, $\theta_A > \theta_B$. We would not know this beforehand, so the goal of the clinical trial will be to compare the performance of device A versus B. For example, we may want to determine whether or not device A has an equivalent inappropriate therapy rate as device B. In this case, we could

set up the null and alternative hypothesis as follows:

$$\begin{cases} H_0 : \theta_A = \theta_B \\ H_1 : \theta_A > \theta_B + \epsilon \end{cases} \quad (3.11)$$

Here, ϵ is the effect size and represents the difference in the inappropriate therapy rates.

For a single trial, we would sample a patient cohort of $x, \{x_i\}$ for $i \in 1 \dots N$ and randomly allocate a sample (patient) to an arm of the trial. Fig. 17 (top) shows one example of a sampled population of size 1000. In our example, the control arm would receive device B and the treatment arm would receive device A. The estimate of $\hat{\theta}_D$ is computed as follows:

$$\hat{\theta}_D = E[Y_D] = \frac{\sum_{i=1}^N Y_D}{N}, \quad (3.12)$$

for each device $D \in \{A, B\}$. The number of inappropriate therapies observed would be counted and depending on the estimate, we would either reject the null hypothesis in favor of the alternative or accept the null hypothesis. This corresponds to one clinical trial and will result in an estimate $\hat{\theta}_D$ for each device. Ideally, we would want to run the trial multiple times and compute multiple estimates of θ_D . As the trial is repeated, this would result in a distribution of the estimates $\hat{\theta}_D$ as shown in Fig. 17 (mid, bottom). As you can see in the figure, the inappropriate therapy rate of Device A is higher than Device B. The degree of separation between the two distributions would depend on the effect size ϵ . However, the reality is that very few clinical trials are repeated multiple times and depending on the distribution of the estimates, it may be more difficult to come to this same conclusion. We explore this concept in the next section.

3.3.4. *Effect of sample size on trial outcomes*

Ideally, in any clinical trial, having a large sample size will be helpful. This is illustrated by Fig. 18. The top panel shows the distribution of $\hat{\theta}_D$ when the sample size is 100. The bottom panel shows the distribution of the estimated $\hat{\theta}_D$ when the sample size is 1000. As

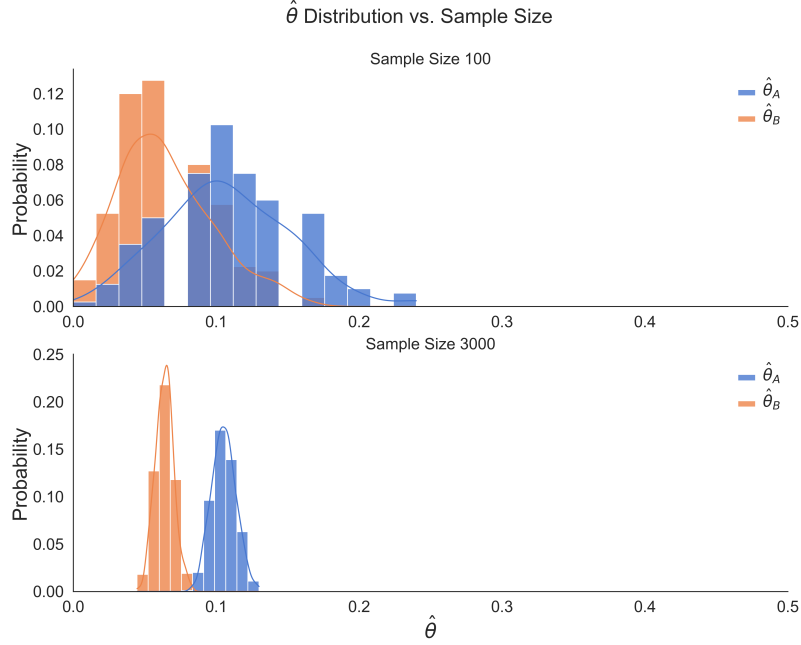


Figure 18: Comparison of sample size and distribution of estimated $\hat{\theta}$. When the effect size is small, the distributions of $\hat{\theta}$ are more overlapped and therefore harder to observe a difference between the two devices with a single trial (top). Even in this case, when the sample size increases, the degree of separation between the distributions of $\hat{\theta}$ also increases.

can be observed, for a fixed effect size, a smaller sample size makes it harder to observe a difference between the two devices. Since we assumed that $\theta_A > \theta_B$, this means that with a single clinical trial the chance that one would derive the incorrect conclusion and reject the null hypothesis is greater. However, the bottom panel shows how even if the effect size is small and it is difficult to determine the difference, with a larger sample size, the distributions have more separation and the relative performance can be distinguished much more easily.

3.3.5. Sample size versus power

In an actual clinical trial, one wouldn't be able to repeat the trial multiple times. Therefore, instead different metrics, such as the statistical power can be used to determine the quality of a trial as explained previously in Sec. 3.2. In our example, since we assume that $\theta_A > \theta_B$, we can estimate the power by fixing the sample size, repeating the trial multiple times, and then computing the proportion that the clinical trial is able to determine the correct

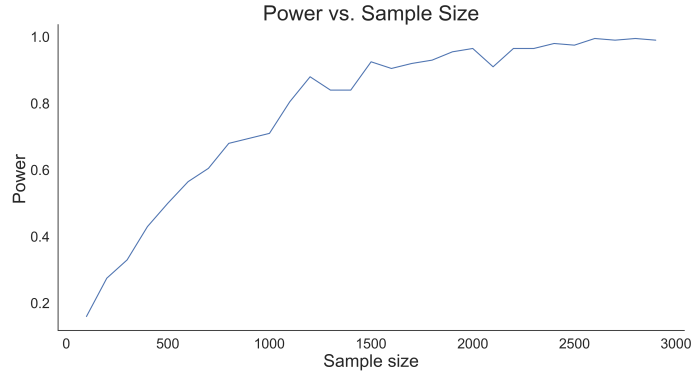


Figure 19: Comparison of sample size and power. Increasing the sample size results in a higher statistical power for a clinical trial. This motivates many approaches for improving clinical trials to focus on the sample size.

outcome and reject the null hypothesis. The relation between the sample size of the trial and the power of the trial is shown in Fig. 19. As can be observed, the power increases as the sample size increases, as expected.

3.3.6. Limitations of increasing sample size

An evaluation with a clinical trial can benefit greatly from a larger sample size. However, simply increasing the sample size is not always possible and may even violate major codes of ethics for clinical trials. Minimizing both risk and inconvenience of participants is of utmost importance when designing and executing a clinical trial (Emanuel, Wendler, and Grady, 2000). Careful preparation during the planning of a trial must be made to ensure that the minimum amount of resources are used to draw a significant conclusion. Various approaches exist in clinical trial design and analysis, but we will focus on approaches to improve the effective sample size in order to increase the power of a trial. Historical controlled trials are one such approach which we describe next.

3.4. Historical Controlled Trials

Historical controlled trials are trials which utilize the data from the control arm of previous clinical trials as depicted in Fig. 3(a). In a pharmaceutical trial, the control arm would often be the response to a placebo. As a concept first introduced by Pocock in the 1970s

(Pocock, 1976), historical controls have the potential to minimize costs and risks to patients by reducing the necessary sample size needed to complete a clinical trial. For example, historical controls are often used in the case of rare diseases which have a narrow window for treatment, but have few too patients to safely execute randomization with placebo controls (Lim et al., 2018).

However, various unique challenges exist in the utilization of historical controls beyond the typical constraints which are needed in general clinical trial design. From a trial design perspective, the differences in the design from trial to trial, such as the duration of the trial, enrollment period, geographic location, etc. can be a major source of heterogeneity in the outcomes (Giorgini et al., 2014; Muntner et al., 2019). Moreover, as time passes the standard of care may change, resulting in a increased placebo response. This makes the effect size observable in a current prospective trial to be reduced. Efforts are underway to promote standardization of measurement procedures and increase harmonization across trials in order to enable comparability (Vesper, Myers, and Miller, 2016). However, we focus more on the patient level variability which can affect the use of historical controls.

From a patient-level perspective, various sources of heterogeneity exist which make the utilization of historical controls difficult. A common source are the differences in demographics as the differences in age and sex can contribute to inconsistent outcomes across clinical trials. For example, experimental placebo responses have shown to be more positive in men (Vambheim and Flaten, 2017). Variability in the baseline severity of a disease is also a cause for greater placebo response. In a clinical trial regarding neuropathic pain, when patients were asked to record a baseline severity of pain, the higher variability led to a higher placebo response (Farrar et al., 2014).

Rather than simply using historical information by itself, combining historical control data and concurrent controls or *dynamic borrowing* has been a more common approach. The *dynamic* part of the term signifies the weight that is placed on historical control data when combining with concurrent controls. Various approaches to dynamic borrowing have been

introduced since Pocock, which attempt to account for such issues and effectively incorporate historical information. These approaches can broadly be categorized into frequentist approaches and Bayesian approaches. For the frequentist approach, the most basic method is *simple pooling*, which makes a strong exchangeability assumption between the historical control data and the current trial. A step beyond this is *testing then pooling*, where a weight of 0 or 1 is decided on the historical data based upon some significance testing for similarity. In other words, according to some measure, if the historical data and the current data is similar to some significance, then the historical data is used, otherwise excluded. However, incorporating historical controls can lead to inflated mean square error, inflated type I error rate, and possibly even lower power. We will explore this last possibility below.

In order to account for the variation that exists in the historical and concurrent control of a prospective trial, propensity-matching (Lin, Gamalo-Siebers, and Tiwari, 2018) is a method that has been introduced. In propensity-matching, the assumption is that the treatment assignment and the outcome are conditionally independent given a set of measured or observed covariates. This allows for differences in populations to be adjusted for according to those covariates by balancing those factors between the population. However, propensity-matching does nothing for unobserved confounding factors which can play a significant role on the outcome. Typically in an RCT, randomization would take care of the bias that can occur due to these unobserved factors. This approach can therefore lead to greater uncertainty in the outcome due to these unobserved factors.

In this thesis, we address the problem of improving statistical power using a Bayesian approach. Bayesian statistical approaches are often better suited to model and account for uncertainty in factors through the use of prior and posterior probabilities, which would not be possible in frequentist approaches. One branch of work utilizes hierarchical models to model the similarity between historical and concurrent control data. For example, a well known method is the meta-analytic-predictive approach (MAP) (Neuenschwander et al., 2010). In this method, between-study variation is estimated using meta-analysis and then

used to estimate the actual outcome parameter of the current trial. The relation between the between-study variation and the outcome is modeled using a hierarchical model. A robust version which include a mixture of the hierarchical model and a vague component in the prior have also been proposed (Schmidli et al., 2014).

An alternative approach taken by this thesis is to incorporate historical data and accounting for the uncertainty by modeling the historical data as a *power prior* (Joseph G Ibrahim et al., 2015a). Unlike Pocock’s method which uses a fixed 0 or 1 for the weighting of historical information, in the power prior method, the likelihood of the historic data is raised to a power α . An α equal to 0 implies a complete downweighting of the historical control and a value of 1 equates using all the historical data. As the historical data is modeled as a prior distribution, the uncertainty in the information and the effects on the overall outcome is made more explicit. Various approaches propose different methods of both defining the power prior or determining the value of the power parameter.

3.4.1. Limitations of historical controlled approaches for clinical trials

Regardless of the approach, historical controls have a fundamental limitation in that the data is fixed and already observed. Even if all the confounding factors are observed and measured, the variability between trial design and patient-level variability may negatively affect the usage of historical controls. In (Schoenfeld et al., 2019), the total sample sizes needed when using historical controls versus concurrent controls in a clinical trial for amyotrophic lateral sclerosis was analyzed. The results showed that between-trial heterogeneity had a strongly negative impact on the power of the trial. Only when there was a large effect size of at least a 35% difference in slopes between the treatment and control was using historical controls more efficient. When the effect size was smaller, the between-trial heterogeneity led to a dramatic increase in the sample size when using a historical control.

3.4.2. Historical controlled trials for the hypothetical device trial

This limitation can be demonstrated in the hypothetical device trial. For the device, assume there are three sets of endpoints with the following distributions:

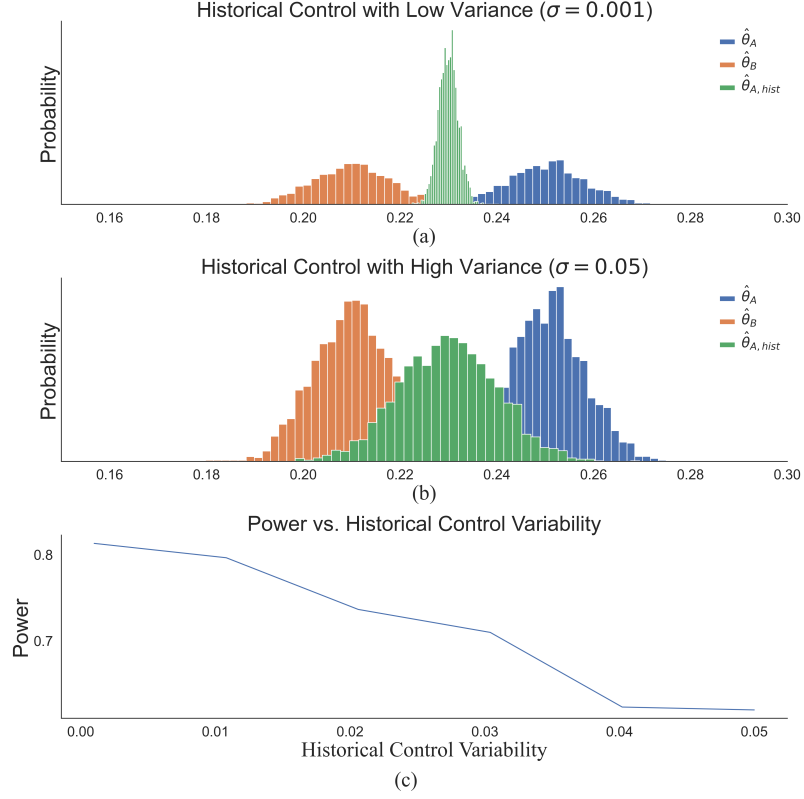


Figure 20: Relation between historical data variability and the power of the trial. When utilizing a historical control in a clinical trial, the estimates of the historical control, $\hat{\theta}_{A,hist}$ can be represented as a Gaussian distribution and the variability can be modeled as the variance parameter σ . (a) Depicts a case when the variance of the historical control distribution is low and (b) depicts a case when the variance is high. As the variability of the historical control increases, the power of the clinical trial decreases.

- For the treatment arm, the endpoints observed from the current population are real, $Y_T \sim f_{Y,real}(y|x, \theta_T)$. The treatment arm receives device B, so $\theta_T = \theta_B$.
- For the control arm, the endpoints are a combination of those receiving device A and data from prior clinical trials for device A. The endpoints for the control arm from the current prospective trial are modeled as, $Y_C \sim f_{Y,real}(y|x, \theta_C)$.
- For the historical data, the conditions may be significantly different from the current prospective trial, so we model it as a different distribution $Y_{hist} \sim f_{Y,hist}(y|x, \theta_{C,hist}, \gamma)$. Here γ is a parameter which determines the variability of the endpoints. For example, if Y_{hist} follows a Gaussian distribution, γ could be the variance parameter, σ .

In a single trial, a sample of the endpoints is drawn for the treatment arm, control arm, and the historical data. The historical endpoints and the control arm endpoints are combined using the test-then-pool method and with the treatment arm endpoints, used to derive the outcome of the hypothesis test as before. As can be seen in Fig. 20(c), as the variability increases, the power of the trial decreases, similar to the conclusion in (Schoenfeld et al., 2019).

From the example, rather than simply using the historical data as is, one can imagine that being able to obtain new data or even possibly generate this data would allow for a source of information more useful than historical controls. This is precisely the approach used in the in-silico evaluation of medical devices, which is covered next.

3.5. In-silico Clinical Trials for Medical Device Evaluation

In comparison to historical controlled trials, in-silico modeling of physiology and simulation allows the generation of new outcomes to evaluate the device. In the case of medical devices, and in particular the ICD, a model of the device and a physiology can be used to simulate and generate a *virtual cohort* of endpoints.

With the hypothetical device, if we assume that an ideal simulator existed which allows for clinical outcomes to be directly simulated, then we would be able to generate data from arbitrary segments of the population. For the medical device $Y = f(x, \theta)$, we can generate samples of the input through simulations of a physiological model, X_{virt} , and generate a virtual cohort of endpoints, Y_{virt} .

As shown in Fig. 21(a), assume that we can generate additional clinical outcomes which can then be combined with actual data within the clinical trial. Note that this has a strong assumption that the exchangeability of the synthetic data and the real data holds. In Fig. 21(c)(d), incorporating the the virtual cohort data has the effect of improving the variability in the estimates of the parameter of interest. In this case, by adding the virtual data we can greatly improve the power, as seen in Fig. 21(b) where the power increases at a

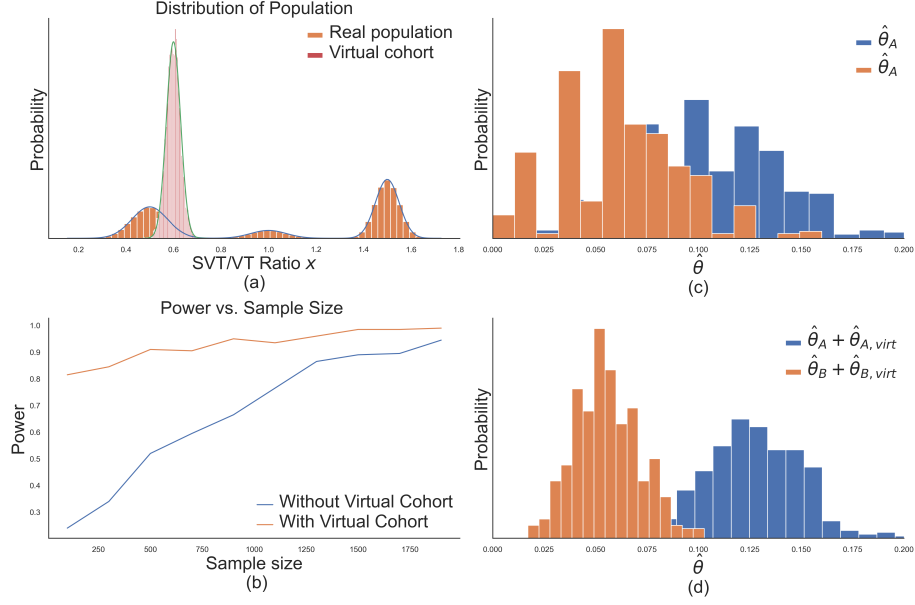


Figure 21: Benefit of an in-silico simulator and in-silico evaluation. (a) An ideal simulator can generate additional data from segments of a target population as a *virtual cohort*. (b) In this case, regardless of the sample size, incorporating the virtual cohort increases the power of the clinical trial. (c) Even when the difference in device performance is not as apparent, (d) incorporating the virtual cohort data increases the separability between distributions as it is equivalent to increasing the sample size.

much higher rate when using the virtual cohort. Incorporating the virtual cohort effectively increases the sample size, thus increasing the power.

3.5.1. Computer modeling and simulation in medicine

Utilizing computer modeling and simulation in medicine and in clinical trials is an active area of research and development in both academia and industry. In pharmaceuticals, pharmacodynamics (PK/PD) models have long been used to study and predict effects of new drugs. The outcomes have been utilized at various stages of clinical trials and an extensive review can be found in (Pappalardo et al., 2018). The issues and challenges with in-silico clinical trials for pharmaceuticals are similar to those of medical devices, but the specific applications are beyond the scope of this thesis.

In addition, clinical trial simulation and modeling has been used extensively in determining components of clinical trials (I. Abbas, 2016). These include testing for statistical model

Table 1: Examples of Physiological Models for In-silico Evaluation of Medical Devices

Reference	Application	Modeling Approach	H/W or S/W?	CT Evaluation?
Simalatsar et al., 2019	Drug delivery	PK/PD	S/W	Possible
Tivay, Kramer, Hahn, 2021	Fluid resuscitation	Whitebox + Blackbox	S/W	Possible
Haddad et al., 2018	ICD lead failure	Bayesian	H/W	Confirmatory
Wilkoff et al., 2013	ICD lead for MRI	Whitebox	H/W	Confirmatory
UVA/Padova	Artificial Pancreas	Whitebox	S/W	Pre-clinical

selection, determining sample size, and also predicting the outcome of clinical trials (Giovagnoli and Zagoraiou, 2012). The work in this thesis is complementary to these methods in that they are more concerned with the higher level design of a trial. It would be possible for the methods proposed in this thesis to be combined with techniques for clinical trial design in order to further improve the efficiency of the design.

The increasing importance of in-silico design and evaluation of medical devices requires widespread collaboration between academia, industry, and regulatory institutions. Efforts to standardize the usage and requirements for computer modeling and simulation include organizations such as Avicenna (Viceconti, Henney, and Morley-Fletcher, 2016) and the Medical Device Innovation Consortium (MDIC). Standards such as the ASME V&V 40 have also been presented to address concerns for rigorous verification and validation of computational models (Pathmanathan and Gray, 2013). In particular, the ASME V&V 40 focuses on the verification and validation of the computational models used in the development of medical devices and is complementary to the methods presented in this thesis.

3.5.2. Notable examples of physiological modeling for in-silico evaluation

Table 1 highlights some notable examples of physiological models developed for the purpose of in-silico evaluation. As shown in the table, depending on the application, various approaches for modeling physiology was used. For example in (Simalatsar et al., 2019), a PK/PD model was used to estimate the response to intravenous anesthetics. In (Tivay, Kramer, and Hahn, 2021), a fluid resuscitation model was developed using a whitebox modeling approach for physiological response combined with a blackbox, variational inference

model for generating parameters for patients. In (Haddad, Himes, and M. Campbell, 2014) and (Wilkoff et al., 2013), a mechanical model of ICD leads was used to predict mechanical lead failure. In (C. Toffanin, M. Messori, F. Di Palma, G. De Nicolao, C. Cobelli, L. Magni, 2014), a simulator for type 1 diabetes mellitus model was presented.

These works can also be compared to whether or not the software component or hardware component of the device was the intended application. This thesis focuses on the software algorithm of the ICD, similar to (Simalatsar et al., 2019; Tivay, Kramer, and Hahn, 2021; C. Toffanin, M. Messori, F. Di Palma, G. De Nicolao, C. Cobelli, L. Magni, 2014).

All of these works demonstrate the potential for utilization as a source of information in a clinical trial of the device. However, only (Haddad, Himes, and M. Campbell, 2014) and (Wilkoff et al., 2013) had an actual clinical trial to confirm the results. However, in this work, simulated evaluations were used separately from real data. In this thesis, we aim to go one step further from these works and utilize the data from physiological models and combine it with real data that would be observed in a clinical trial. This is possible due to the unique aspects of in-silico evaluation of medical devices, which we discuss next.

3.5.3. Unique considerations for in-silico evaluation medical devices

Medical devices are unique from pharmaceuticals in that they provide greater opportunities for utilizing in-silico evaluation (Faris and Shuren, 2017). First, for many medical devices the most important aspects of safety are not even related to a clinical trial. Therefore, extensive lab benchmarks and simulation play a major role in establishing the evidence for safety and clinical trials often act as a confirmatory process over a limited subset of device operating conditions. For example, with the Medtronic Revo MRI pacemaker system, establishing that patients implanted with the device would not be subject to damaging levels of heat produced when the patient undergoes an MRI scan (Kalin and Stanton, 2005). Wilkoff et al. (2013) utilized computer simulations to test 2.4 million combinations of patient anatomy, MRI coil, position in the MRI coil, and lead path. Additional validation of results were studied in vivo and with data collected through animal experiments. A subsequent clinical

study was able to confirm the results of the analysis and supported the expected safe performance (Gold, Sommer, et al., 2015).

Another unique aspect medical device evaluation is that an ample body of evidence can be found from clinical experience. This experience can be combined with extensive knowledge regarding the physiology in order to create a virtual patient model. For example, the Type 1 Diabetes Mellitus Padova/UVA simulator (C. Toffanin, M. Messori, F. Di Palma, G. De Nicolao, C. Cobelli, L. Magni, 2014) is another notable example of utilizing in-silico evaluation for determining the safety of medical devices. In this case, the model was used for preclinical safety assessment of the control algorithms of artificial pancreas technologies. After extensive validation, the model was accepted by the FDA as an alternative to animal testing. This work is similar to the results presented in this thesis in that the purpose of the model was to test software control algorithms. However, this focuses on utilizing the information generated from the simulator in the context of a clinical trial and within the domain of cardiac modeling.

In (Haddad, Himes, and M. Campbell, 2014), the authors developed a statistical cardiac lead fracture model which could be used as a virtual patient model to evaluate lead fatigue survival. The authors were able to compare the fracture survival in the virtual patient models to the observed field performance with comparable results. A Bayesian statistical approach was utilized to build the virtual patient model. However, the evaluation within clinical trials did not explicitly model potential sources of uncertainty which can affect the outcomes. We motivate the importance of considering sources of uncertainty in the next section.

3.5.4. Limitations of in-silico clinical trials and motivation

Up until now, we have assumed that a simulator would be ideal and the outcomes in an actual clinical trial can be estimated predicted accurately. The reality, however, is that most physiological simulators are not ideal and inherently have inaccuracies which may or may not be measurable. These inaccuracies can have significant effects on the outcomes.

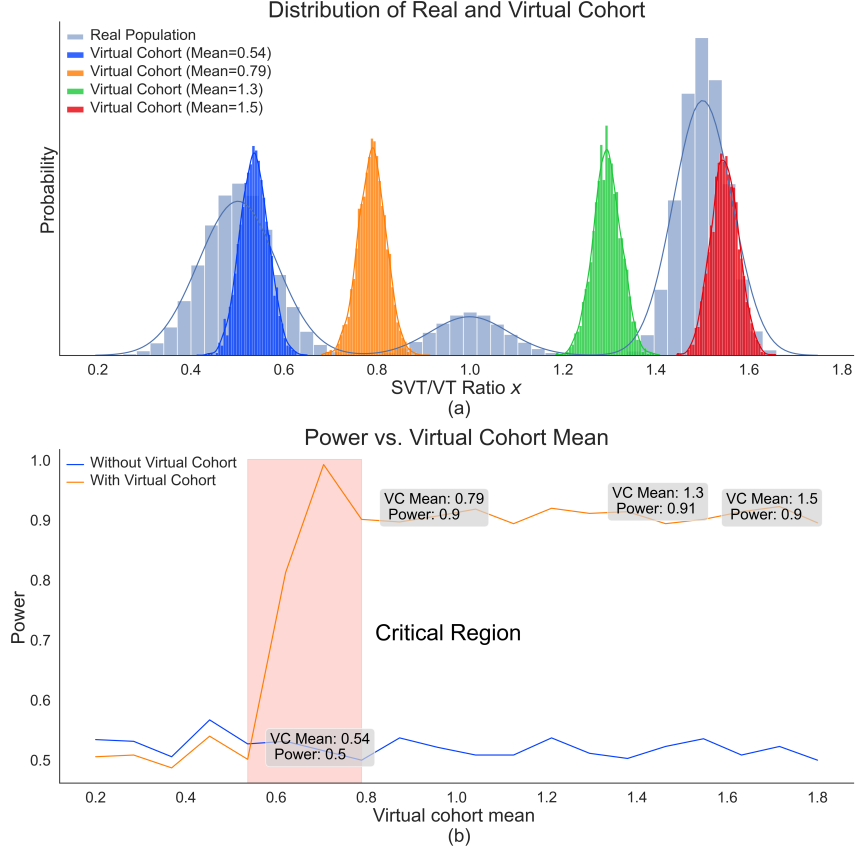


Figure 22: Effects of simulation uncertainty on the outcome of the trial. (a) Virtual cohorts with different means represent simulation uncertainty. (b) When the virtual cohort is generated near a critical region of the target population, the power of the clinical trial can change drastically depending on the mean of the virtual cohort. Outside the range, the effect on the power is less pronounced and the uncertainty in the outcome is reduced.

For the hypothetical device, an example of such inaccuracy and the effect of the outcome is depicted in Fig. 22. We can model the inaccuracy of the device simulator by limiting the range of the population from which the simulator can sample. In our example, we limit the variance of the virtual cohort and generate samples for different means of x . This is shown in Fig. 21(a) by the virtual cohorts with different means. In Fig. 22(b), if the range that the simulator can sample from falls near a region of the population where the differences of the device performance is most pronounced, then combining the outcomes from the simulator with the real data will improve the clinical trial and improve the power. However, if the range of the simulator falls on the wrong side of the critical range, then it's possible that the simulator can in fact bias the outcome in a negative direction and degrade

the power.

Another consideration is that if the simulator range is at the boundary of the critical region, then it is possible that slight changes in the simulator range will result in differing outcomes. For example, in 22(b), near the critical region changing the range of the simulator slightly can result in a drastic change in the power of the trial. If this was the case for an actual device, regardless of the outcome of the trial, one would have to doubt the certainty of the conclusions derived from the clinical trial. In comparison, in 22(b), outside of this critical region, slight changes in the simulator have little or no effect on the power of the clinical trial. This adds to the confidence of the outcomes derived by adding simulation information. In either case, quantifying the degree of uncertainty would aid in interpreting the results.

In this hypothetical device trial, we know the exact thresholds and which ranges of the simulator can affect the outcome. However, in an actual clinical trial and with a real simulator, these factors would not be straightforward and would require some framework of analysis in order to consider this source of uncertainty. This issue and the limitations of the related works motivate the main approach of this thesis and key questions, which we describe in the next section.

3.6. Main Approach and Key Questions

We can visualize the various approaches for improving the power of a clinical trial according to the additional source of information and its characteristics as shown in Fig. 23. First, we can think in terms of the validity or how representative the data is with respect to the measurement of performance. Second, we can think about it in terms of the quantity of data which is related to the cost of producing the data. If the data is easy to produce and does not incur a lot of cost, then the quantity can easily be increased. In terms of validity, the randomized control trial has the most validity due to the randomization and the fact that real humans participate in the trial. However, the costs are enormous, hence reducing the quantity that is available.

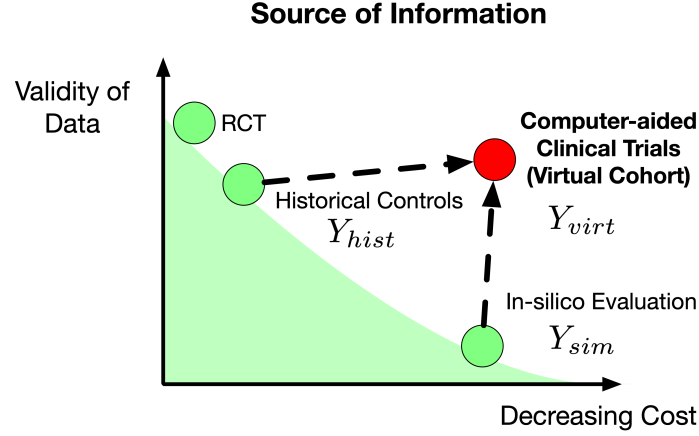


Figure 23: Comparison of methods by sources of information. At one extreme are RCTs with historical controls being an improvement in terms of cost. At the other extreme is in-silico evaluation data, which is inexpensive to obtain, but the validity is of concern. Our approach, the Computer-aided Clinical Trial (CACT) aims to incorporate the advantages of both historical controlled trials and in-silico evaluation to improve the power of a clinical trial.

With historical controls, the quantity of information available is more abundant than RCTs. However, as the data is fixed it is less valid as there is a potential for bias in the outcome. Moreover, at some point in time a cost was incurred to execute the clinical trial.

In comparison, in-silico evaluation is very flexible in that you can easily simulate a large amount of data as long as a simulator exists. However, the validity of the data comes into question, and as observed in the hypothetical example, the uncertainty in the simulated data can potentially have drastic effects on the outcomes.

Based off these observations, in this thesis, we propose an approach that aims to be the best of both worlds and combine techniques from both historical controlled trials and in-silico medical device evaluation. In terms of quantity we wish to combine synthetic virtual cohorts with real data, thereby reducing the need for real data. Additionally, we want to quantify the uncertainty in order to improve the validity of the data. In order to achieve this goal, we resolve the following key questions in order to develop our main approach:

1. How do we generate a simulated, virtual cohort of data for device evaluation while systematically incorporating historical information?

2. How do we combine the virtual cohort with real data observed in a clinical trial in order to improve the power of the trial?
3. How do we account for the uncertainty in the synthetic data when drawing conclusions for the clinical trial and quantify such uncertainty?

In this thesis, we address these first three questions using a Bayesian statistical approach and develop a framework we call the *Computer-aided Clinical Trial* (CACT) for medical devices. Specifically, we incorporate the cohort of synthetic data, which we call a *virtual cohort*, by combining the power prior method (Joseph G. Ibrahim et al., 2015b) and building a framework of evaluation which can model sources of uncertainty. With this framework, we are able to define a measure of uncertainty called δ -robustness and use it to analyze and interpret the outcomes of a CACT.

The first three questions are related to the evaluation of the device and improving the evaluation process with a clinical trial. However, the evaluation does not improve the actual performance of the device. Therefore, this thesis also examines the following question:

4. How can we improve device performance with an automated, data-driven approach?

3.7. Chapter Conclusion

In this chapter, we first formalized the definitions of a clinical trial and power which are the main concerns of this thesis. We introduced the trade-off between the sample size and the power of a clinical trial through a hypothetical medical device and target population. We introduced historical controlled trials as an approach for improving the power of trial with limited sample size and the limitations of such approaches. As a parallel effort, in-silico simulation has been considered as an alternative source of information to be used in the context of clinical trials. We presented the recent developments and initiatives regarding in-silico evaluation and demonstrated the limitations of the approach. In the following chapters, we begin to address the questions presented in this chapter within the context of the ICD and a clinical trial for evaluating ICD, the RIGHT trial.

CHAPTER 4 : Virtual Cohort Generation and In-silico Medical Device Evaluation

4.1. Overview

In this chapter, we describe in detail the process of generating a cohort of synthetic physiological signals using a simulator. Elements of this chapter have been adapted from “High-level modeling for computer-aided clinical trials of medical devices” in the Proceedings of the High Level Design Validation and Test Workshop 2016 and **jiang'-silico'2016** in the Proceedings of the Conference of the Engineering in Medicine and Biology Society 2016. These papers were joint work with Houssam Abbas, Zhihao Jiang, Marco Becanni, Jackson Liang, Sanjay Dixit, and Rahul Mangharam.

4.2. Introduction

Medical device verification with clinical trials. During a typical hardware and software development process, verification activities take up a major portion of the process. Each verification activity has a sign-off criterion in the verification test plan, indicating that the design can proceed to the next phase of the development cycle. Such criteria will usually include successful linting and sufficiently high coverage metrics (code, functional, data and assertion coverage in particular). Model checking of certain sub-systems is also performed, where the sub-systems are chosen based on their criticality and their size. Finally, integration testing is performed when the system is assembled together.

During this process, verification seeks to determine whether a design has satisfied some specification. This specification is defined according to the requirements necessary for the system to operate properly and safely in the desired environment. For medical devices, especially medical cyber-physical systems such as the ICD, the environment is human physiology. As we observed in Chapter 2, certain aspects of the device can be tested through lab testing and on an abstracted model with verification. However, how a system will perform under large patient heterogeneity cannot be evaluated without a clinical trial. Thus, in some sense, the de facto verification of a medical device requires clinical trials.

As described in Chapter 3, clinical trials are subject to their own set of restrictions and limitations. Therefore, in the remainder of this chapter, we first outline the requirements necessary for providing an additional source of information through simulation of physiological models in the context of a clinical trial. Next, in line with these requirements we introduce the synthetic electrogram (EGM) generator used to generate a virtual cohort of physiological signals necessary for the evaluation of an ICD discrimination algorithm. With this cohort, we apply the generated virtual cohort to evaluate the ICD discrimination algorithm and demonstrate the ability to predict the outcomes of an actual clinical trial for ICDs. Finally, we improve the virtual cohort generation process by applying Bayesian hierarchical models and demonstrate the improvement in prediction.

4.3. Virtual Cohort Generation

4.3.1. Necessity of physiological simulation

From a cursory understanding of the problem, it is reasonable to question the necessity of physiological simulation rather than simply utilizing a dataset of pre-recorded signals. We address this concern with the following reasons for pursuing a physiological simulator:

- (Taming input complexity) The space of physiological input signals is complicated and with no evident structure. Formally, the input space is uncountably infinite since the physiological signal is real-valued, unlike the input to, say, a network router which contains discrete values. By modeling, we obtain a finite representation of the input space, impose a structure on it, and obtain a test of what is a physiologically valid signal (one that can be produced by the model) and what is not.
- (Separation of design and validation) Medical device companies likely have access to a vast set of data that is retrieved from their devices. This data might then be used to develop and test the device. Because a clinical trial is meant to be an independent assessment of the device’s performance, the data used to develop it cannot be re-used in a clinical trial. A clinical trial that re-uses the development data will likely show

very good performance and bias our estimate of true performance.

- (Paucity of data) Physiological data is not readily available. By physiological data we mean the signals that are measured by a medical device and which it uses to diagnose the state of the patient and apply appropriate therapy. What data is available is usually siloed in proprietary platforms.

4.3.2. Requirements of a physiological simulator

Given the previous argument, in order to understand the general requirements of a physiological simulator, recall the generic medical device which was defined in Chapter 3, A medical device is a function which takes as input the data X and a parameter(s) of interest θ and outputs a decision Y :

$$Y = f(x, \theta). \tag{4.1}$$

It is well-established that the higher the level of the abstraction, the easier it is to design system inputs and the easier it is to interpret test results. Moreover, the outputs of our hypothetical device, Y , are used as endpoints to determine the outcome of a clinical trial.

Therefore, we wish to obtain the highest level of abstraction possible for the input X which would allow us to reliably reflect the underlying behavior of a patient, while still interfacing with an actual implementation of the device and allow a proper input-output relation. In other words, not only do we want to be able to interface with the device at a highest level possible, but still be able to reliably control the necessary characteristics of the generated data for our intended purposes.

The specifics of a physiological model would highly depend on the system that is being evaluated and details of the specific clinical trial, in particular the endpoints of the trial. For example, in the case of the ICD, the main aspect that we wish to evaluate is the discrimination algorithm. In the RIGHT trial (Gold, Ahmad, et al., 2012), the time-to-

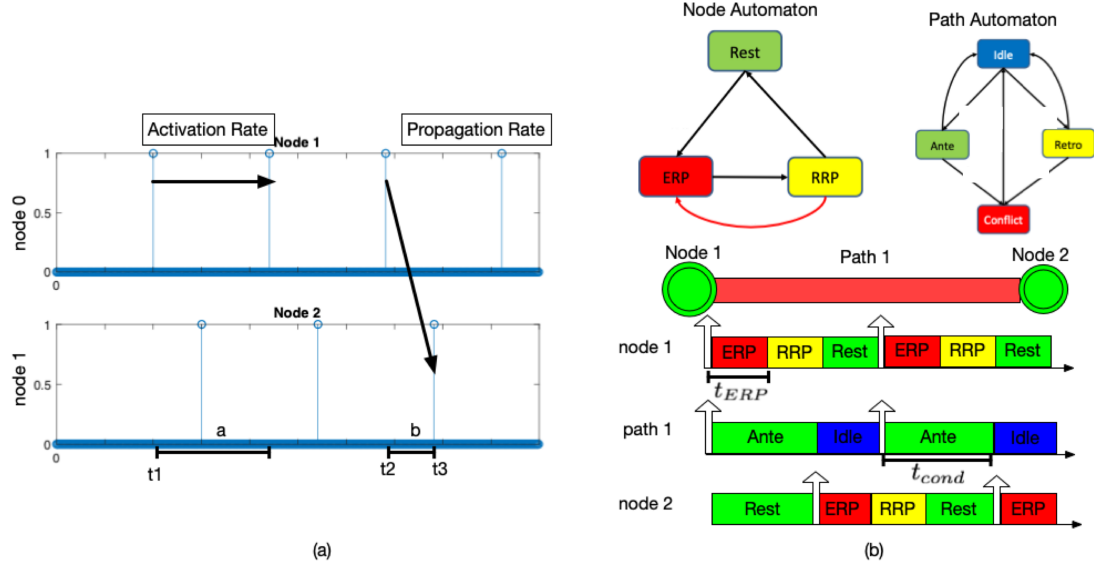


Figure 24: (a) An excerpt of a simulation trace for the TA model, the extracted features, and the node and path automata. The activation rate is measured as the time between consecutive activation of a node (t_1) and the propagation rate is measured as the time between activation of a node and the activation of the neighboring node ($t_3 - t_2$). (b) The TA model is composed of node automata and path automata. A node represents the electrophysiology within a specific region of tissue and a path captures the conduction between two regions.

first inappropriate therapy was the primary endpoint. This requires the ability to simulate instances of appropriate or inappropriate therapy by the device. Since the input to the device algorithm is an EGM, we first need to be able to generate a synthetic EGM. The algorithm operates based on the timing and morphology of EGMs. Therefore, modeling these characteristics will allow for the generation of the necessary synthetic EGMs. In the subsequent sections, we present a timed-automata based EGM model which we use throughout the remainder of this thesis as the generator of synthetic EGMs.

4.3.3. Timed-automata based EGM simulator

EGM Timing Model. In order to simulate the timing behavior, we utilized a timed-automata (TA) based formalism to model the behavior as presented in (Z. Jiang, Pajic, Moarref, et al., 2012). The TA model abstracts behavior of the heart, capturing only activation and conduction through the tissue. This facilitates testing of devices which do not rely on a high-fidelity signal such as that produced by the Fenton-Karma model

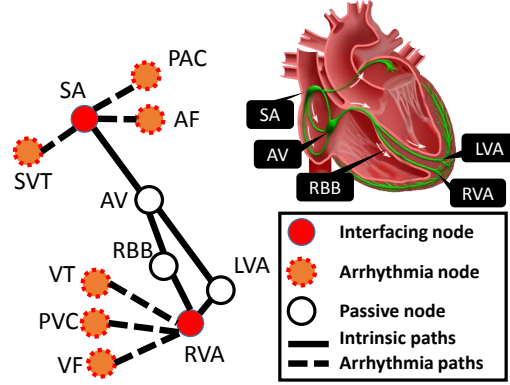


Figure 25: TA model of electrical conduction system of the heart. Each component is represented either as a node automata or path automata. In addition to the main conduction pathway from the SA node to the right ventricular apex (RVA), nodes representing the sources of various arrhythmia are also included in the model.

(F. Fenton and Karma, 1998). The reasoning is that modeling individual cells in order to obtain a global view of the heart is processor heavy and contains extraneous information for purposes of cardiac device testing, which does not rely on such a high-fidelity signal. Instead, the model utilizes the timing properties of the heart to obtain a macro-level view by lumping cells into a node automaton and path automaton. The model captures the timing behavior of the conduction network of the heart and various cardiac tissue by lumping cells into a node automaton and path automaton. The key observations that allow for such an abstraction are that the activation of a tissue only occurs locally among neighboring tissue and once a section of cardiac tissue has been activated and depolarized, it cannot be activated for a brief interval known as the refractory period (see Sec. 2.3).

A section of tissue is represented as a state machine defined as a composition of two node automata and the conduction pathway between them as a path automaton. The spatial configuration of the different regions of the heart are represented through the topology. This is depicted in Fig. 25. The activation and refractory properties of tissue are represented by the nodes and the conduction properties between the nodes are modeled by the path. As shown in Fig. 25, each solid circle represents a specific anatomical location of the heart. For each location, a node automata models the timing behaviors of the generation and blocking

of electrical depolarizations. In this case, we only model the anatomical locations that can affect the electrical behaviors between the atrial lead and the ventricular lead of the ICD. For each solid line, a path automata models the timing delay between two locations. We also model different sources of abnormal electrical depolarizations and their connections with the main model structure, which are represented as dotted circles and lines in Fig. 25.

$$(O) = f_{node}(k, I) \quad \text{For node automaton} \quad (4.2)$$

$$(O_{ante}, O_{retro}) = f_{path}(k, I_{ante}, I_{retro}; \eta) \quad \text{For path automaton} \quad (4.3)$$

Here O represents the output signal of the automata and I represents the input signal to the automata and k is the index as a particular model structure contains multiple node and path automata connected in a graph. Additionally, η is the set of parameters for the node and path automata. The value of this η determines the behavior of the model and setting η will allow the model to simulate a particular heart condition.

The basic state transitions of the node automata and the path automata are shown in Fig. 24(b). A node begins in an idle **rest** state and activates at the end of its resting period T_{rest} . An activation on one node begins conduction through any attached path automaton. After a finite interval of T_{cond} , the signal is conducted to a node connected to the opposite end of the path and triggers an activation. A node that has been activated cannot be activated again during an absolute refractory period of T_{ERP} , after which the node transitions to the rest state. In a node automaton, the refractoriness is modeled as ERP, RRP and Rest states, and their durations are modeled by the timers T_{ERP} , T_{RRP} , and T_{REST} . In a path automaton, the conduction properties are modeled as no conduction (Idle), antegrade conduction or forward conduction (Ante), retrograde conduction or backward conduction (Retro), dual direction conduction (Double), conflict (Conflict) state. The conduction delays are modeled by the timers T_{ante} and T_{retro} .

EGM Morphology Model. The morphology of the EGM was first determined by extracting condition templates from actual recordings of EGMs in the Ann-Arbor Electrogram

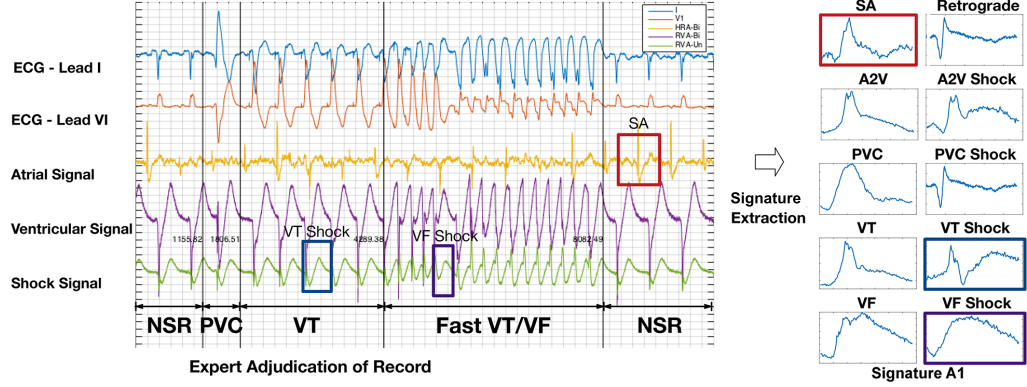


Figure 26: EGM morphologies are identified and extracted from patient episodes. 10 EGM morphologies corresponding to different signal sources are extracted as EGM signature for each patient.

Library (AAEL, J. M. Jenkins and R. E. Jenkins, 2003). A total of 10 different types morphologies were extracted from the data. If a patient was missing a particular morphology, a random selection for the missing morphology was taken from a different patient.

To incorporate beat-to-beat variation, a wavelet transform was applied to an instance of the morphology where the latter stage coefficients were randomly varied.

4.3.4. Simulation of EGM signals

This model structure generates the timing of electrical depolarizations for both atrial and ventricular channels. According to the source of the electrical depolarizations, an EGM morphology is overlayed on top of the timing events, which completes EGM generation. The EGM morphologies are collected from EGM signals of real patients. The EGM generation process is demonstrated in Fig. 27. For a given heart model instance, the parameters of the heart model are set to simulate a particular condition, such as normal sinus rhythm or ventricular tachycardia. The simulation of the heart model results in the event timings for the corresponding channels of the EGM input signal. For the current ICD discrimination algorithm, we simulated the atrial, ventricular, and shock channel signal. The atrial and ventricular channel measure bipolar EGMs whereas the shock channel signal is a unipolar EGM. The event timings are overlayed with the base morphology signatures depending on the type of event that occurred. To account for the beat-to-beat variability in the signal,

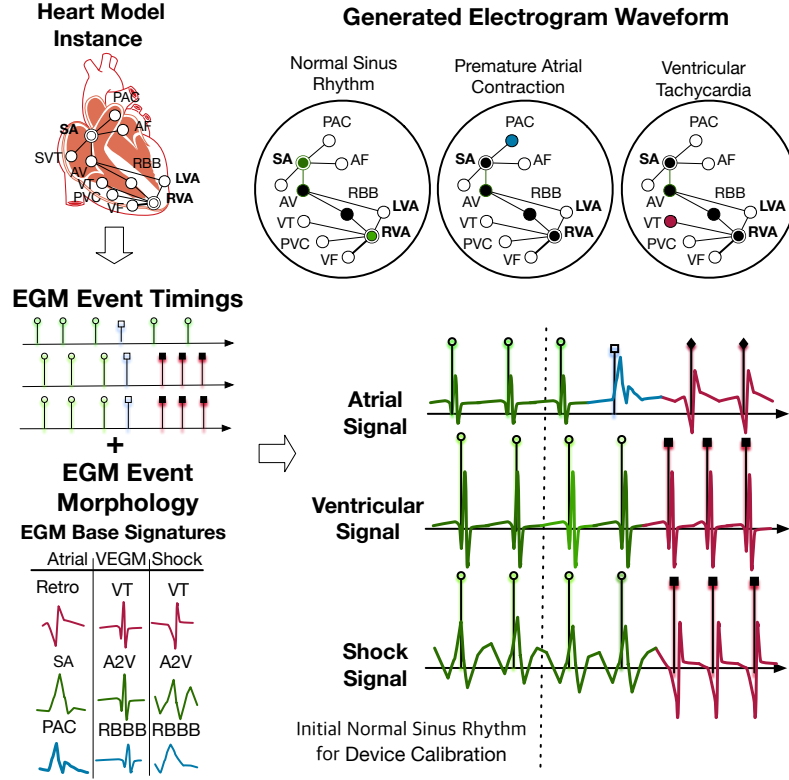


Figure 27: EGM generation process. The timing model produces the events as a boolean signal (top left). The base EGM signatures are overlaid on the events according to the type of event (bottom left) to produce the final signal (bottom right). Different arrhythmias are modeled by different timing models (top right).

the coefficients of the wavelet transform of the signature are randomly perturbed. The event timings combined with the morphology generates the synthetic EGM. The label of the synthetic EGM is automatically generated by construction.

4.3.5. Validation of signals

In order to facilitate the validation of the generated signals with the actual physical device, a hardware interface was developed as shown in Fig. 28. The interface inputs physiological signals to the device and records the response automatically. Scenarios combining several types of rhythms can be programmed and applied to the device.

The expected response of the device and the actual response of the device was compared and validated for a small subset of generated signals.

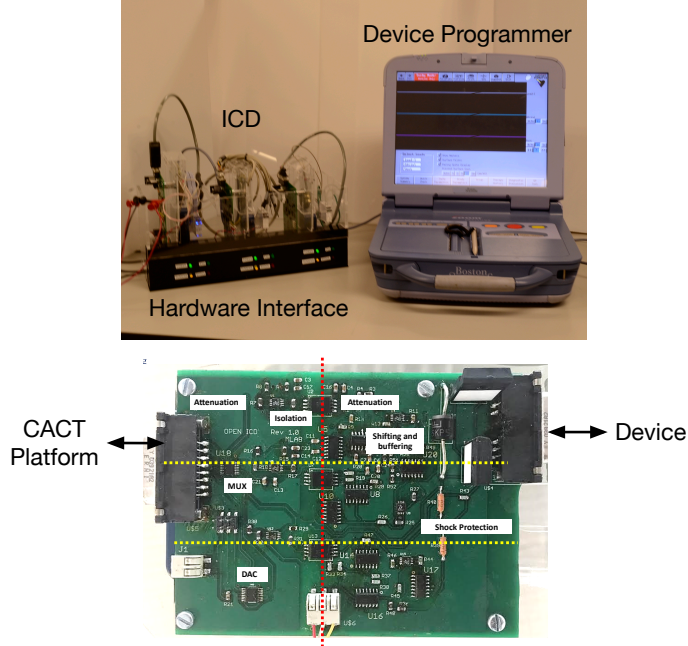


Figure 28: Hardware interface for in-silico evaluation of ICD. Episodes from physiological cohort can be applied to the physical device. The interface can also be used for device model validation

4.3.6. Virtual cohort generation

For each heart condition, we obtained physiological ranges for the timing parameters from the clinical literature (Josephson, 2008), and built a synthetic cohort by uniformly sampling from the ranges. We were able to construct 600 synthetic heart models for each of 19 heart conditions, and simulated the heart models to obtain 11,000+ arrhythmia episodes. These signals are then applied to implementations of ICD discrimination algorithms in order to obtain a virtual cohort.

4.4. In-silico Evaluation of the ICD

With the virtual cohort defined in the prior section, we evaluated a device algorithm according to the objective of a clinical trial. These concepts will be further formalized in Ch. 5, but here we demonstrate the utility of virtual cohort generation for the evaluation of ICD performance in the context of RIGHT.

ID of V2 devices contrary to what was assumed by the RIGHT investigators. Because an ISPCT is designed and conducted in support of a given RCT, the details necessarily depend on the RCT we consider. In Fig. 29 we give an overview of the process of the ISPCT we conducted in support of RIGHT.

- ① Modeling starts by adjudication of 380 individual episodes from a database of 123 EGM records of real patients. Each such episode provides EGM morphologies that are annotated with the tachycardia that produced them (see Fig. 26 for examples), resulting in the identification of 19 different rhythms.
- ② An automata-based timing model is used to simulate the timing characteristics of various tachycardias. Combined with the annotated EGM morphologies, we can now generate parameterized probabilistic heart models that simulate different tachycardias, and variations on each tachycardia.
- ③ A cohort of $> 11,000$ models is generated by varying the parameters of the heart model. The parameter ranges depend on the tachycardia being simulated.
- ④ Every member of the cohort is then simulated to produce EGM signals that are fed to both ICD algorithms. The inappropriate therapy rate of the two algorithms are analyzed.

4.4.3. Results of in-silico preclinical evaluation of ICDs

Estimates of inappropriate therapy rate

The first objective of the ISPCT is to estimate the rate of inappropriate detection for both algorithms across all arrhythmias combined, i.e., for the entire synthetic cohort. The rate of inappropriate therapy is defined as

$$\hat{\theta} = \frac{\text{Number of inappropriately applied therapies}}{\text{Number of applied therapies}}$$

From this, we can validate the assumption that V2 outperforms MDT.

Conclusion 1: MDT (w/ PRL+W) delivers less inappropriate therapy. The obtained rates of inappropriate therapy were 6.65% for V2 and 2.91% for MDT ($P < 0.0001$), assuming an equal number of patients from each arrhythmia in the synthetic cohort. The corresponding relative improvement *of MDT over V2* is 56%. Our findings are consistent with the observations of the RIGHT trial itself (Gold, Ahmad, et al., 2012), and are purely simulation-based.

Conclusion 2: Result holds across population characteristics. The above rates were obtained under the assumption that each arrhythmia is equally represented in the cohort. A significant feature of in-silico trials is that they allow us to study the endpoint of interest (here, rate of inappropriate detection) on a variety of populations, which have the various arrhythmias in different proportions. This may not be feasible in a real clinical trial, which has to contend with the population present at the clinical centers where the trial is conducted. We may then ask: *does MDT maintain a lower rate of inappropriate detection across different populations?* To answer this question, we varied the distribution of the arrhythmias in the synthetic cohort, and re-computed the cohort-wide rates of inappropriate therapy. We conducted trials for 100 random variations of the arrhythmia distribution. Fig. 30 shows the results for the uniform distribution and a distribution that approximates that of RIGHT’s cohort (Gold, Ahmad, et al., 2012, Table 1). It can be seen that indeed,MDT maintains a better rate of arrhythmia discrimination.

These results illustrate the benefit that an ISPCT can bring to the planning of a clinical trial (CT): the fact that V2 could not be shown to be better than MDT can cause the investigators to re-consider their assumptions and the feasibility of the trial. In this case, the ISPCT casts doubt on the assumed *direction* of the effect, i.e. whether the treatment intervention (V2 w/ Rhythm ID) is better than control (MDT w/ PRL+W), or vice versa. This early-stage check can mean the difference between an expensive trial that fails at showing the desired effect, and a successful trial that is appropriately sized.

Thus, while an ISPCT does not replace or mimic the CT, it can provide *early insight* at a

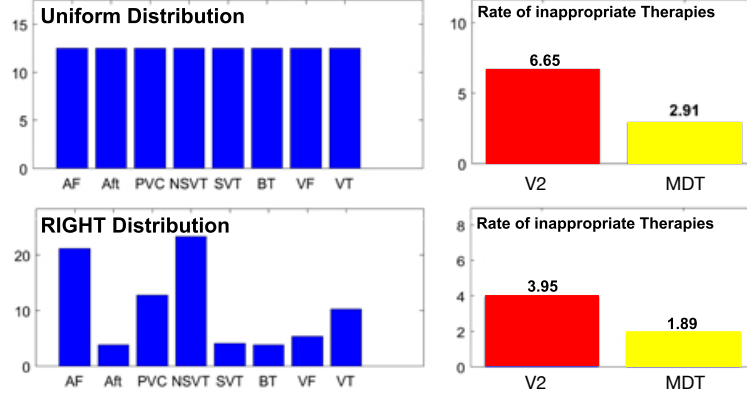


Figure 30: Estimated rates of inappropriate detection (2nd column) for different arrhythmia distributions (1st column). The upper-left distribution is uniform, and the lower-left distribution is that of the baseline characterization in RIGHT (Gold, Ahmad, et al., 2012). (Figure credit: Zhihao Jiang)

small fraction of the CT cost and duration and without violating any ethical issues.

Condition-level analysis

In addition to the estimates of inappropriate therapy rate, the ISPCT allows us to estimate the *sensitivity* and *specificity* of the diagnostic algorithms' performance, something which is not possible in a clinical trial because the device only records a limited number of episodes for which therapy was delivered. These are defined as

$$\text{Sensitivity} = \frac{\text{Nb of correctly detected sustained VTs/VFs}}{\text{Nb of sustained VTs/VFs}}$$

$$\text{Specificity} = \frac{\text{Nb of correctly detected SVTs/non-sus VTs}}{\text{Nb of SVTs/non-sustained VTs}}$$

In words, the sensitivity measures how well the device recognizes sustained VTs. Specificity measures how well the algorithm discriminates between VT and SVT. An ideal algorithm would have 100% sensitivity and specificity. Unfortunately, these are typically competing goals: the more sensitive the device, the more likely it will misdiagnose some SVTs as VTs, so its specificity will drop.

We calculated sensitivity and specificity in our in-silico trial, and report them in Table 2 on a per-arrhythmia basis. The conditions are drawn from RIGHT's baseline characteriza-

Table 2: Specificity and sensitivity of ICD VT/SVT discrimination algorithms

Arrhythmia	Specificity (%)		p-value
	V2	MDT	
Atrial Fibrillation	99.8	99.6	0.3167
Atrial flutter	58.3	79.33	<0.0001
Premature ventricular complexes	100	100	1
Nonsustained ventricular tachycardia	100	99.8	0.3171
Other Supraventricular tachycardia	96.3	99.7	<0.0001
Brady-Tachy	100	98.83	0.0079
	Sensitivity (%)		p-value
Ventricular fibrillation	100	100	<0.001
Ventricular tachycardia	100	100	<0.001

tion (Gold, Ahmad, et al., 2012). It can be seen from these results that in our synthetic cohort, Atrial flutter and other SVTs are the main source of inappropriate detection for V2 compared to MDT. In the case of atrial flutter, V2 categorizes it inappropriately as VT for 41.7% of the episodes.

Condition-level analysis pinpoints the specific decision pathways of the discrimination algorithm which must be addressed to reduce the device’s rate of inappropriate therapy. It is difficult to get such insight through a CT as the patient population is fixed and the conditions are determined retroactively. Such analysis can be further used to investigate condition distributions across different patient populations (e.g. abnormal heart rhythms in children vs geographic region-specific or race-specific condition distributions).

4.4.4. Effect of device parameters on discrimination performance

ICDs have a number of parameters which can be tuned by the physicians to accommodate specific patient conditions. Currently there are few clinical results on the effect of different parameter settings on sensitivity and specificity. One of the main causes of VT/SVT misclassification has been inappropriate parameter setting (Daubert et al., 2008). For the physicians to set appropriate parameters, it is very important to understand how the change of one parameter can affect the discriminating capability of the algorithm. With an ISPCT,

one can subject the same synthetic population to different settings of the parameters at virtually no cost and no risk to patients.

In this section, we use the ISPCT to demonstrate the effects of changing two common parameters on SVT/VT discrimination specificity. The first parameter is the ***VT duration*** of arrhythmia before the ICD makes a therapy decision. The parameter for MDT is the number of consecutive fast ventricular intervals which can be set from 8 to 20 beats. In this experiment we explore the values {8,10,12,16,18,24,30}. From the results (Fig. 31) we observe that the specificity increases monotonically with the length of the duration, which matches the intuition as the device can examine a longer history of the arrhythmia episode with longer duration, and also allows a greater chance for the arrhythmia to self-terminate. This can prevent inappropriate detection therefore prevent inappropriate therapies. However, setting the duration too long can delay, and in some cases withhold appropriate therapy, as sensitivities dropped below 100% when the number of consecutive beats is more than 18.

The second parameter we varied is the ***VF threshold***. If the ventricular rate is faster than the VF threshold for a period of time the algorithm will confirm detection without going into the SVT/VT discrimination algorithm. In this experiment we explored the values {170,184,200} msec. As the parameter increases from 170BPM to 184BPM, more episodes will be examined by the SVT/VT discrimination algorithm, increasing specificity (Fig. 31).

This specific outcome presents an interesting usage of in-silico evaluation for improving the settings of the device algorithm parameters. We further explore this problem in Chapter 6.

4.5. Bayesian Hierarchical Models for Virtual Cohort Generation

Despite the ability to predict the outcome of RIGHT, the results obtained from the proposed model and generated virtual cohort failed to converge to the true rates of the clinical trial. One cause of this is the uncertainty and bias in the generated EGM parameters which are selected during the virtual cohort generation process. This inaccuracy can be modeled as

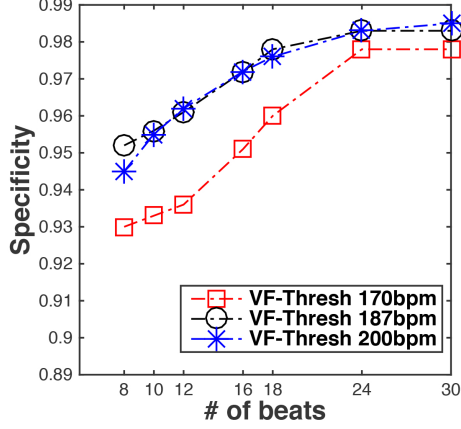


Figure 31: Effect of Duration and VF threshold parameters on specificity. As we increase the duration and VF threshold parameters, we can observe an increase in specificity. This is due to the algorithm having a longer history to examine as well as increases in the chances of self-termination of a VT episode.

a source of uncertainty in the virtual cohort generation process. In the following section, we focus on how Bayesian hierarchical models can be used to systematically incorporate historical information in order to account for such bias.

4.5.1. Modeling uncertainty in virtual cohort generation

In general, the performance of the device will heavily depend on the overall characteristics of various prognostic factors in the synthetic cohort. In the case of the ICD, these factors can include the distribution of the occurrence of various arrhythmia types and the VCL, the peak-to-peak interval of the ventricular EGM.

We can assume that most physiological models will be governed by physiological parameters. Thus, uncertainty in underlying physiological parameters manifests as uncertainty in physiological signals, which will then manifest as uncertainty in the outcomes of evaluation.

We model the uncertainty in the virtual cohort generation process using a Bayesian hierarchical model. In this model we relate the signals generated in a virtual cohort, X_j , to the physiological parameters or *settings* of the physiological model, η , denoted as

$$X_j \sim p_X(x_j | \eta). \quad (4.4)$$

for $j = 1 \dots N_o$, where N_o is the size of the cohort.

For the case of the EGM generator, we denote the uncertainty in the parameter η using a prior distribution, parameterized by condition-specific information, ψ_t , where t denotes the heart condition, such that

$$\eta \sim p_\eta(\eta \mid \psi_t). \quad (4.5)$$

For example, in RIGHT, the physiological signals generated from the model are the EGMs and the condition-specific available information is the VCL for each type of heart condition, $t \in \{\text{NSR}, \text{SVT}, \text{VT}\}$.

4.5.2. Virtual cohort generation for RIGHT with prior distribution for model parameters

In the case of RIGHT, the uncertainty in the EGM stems from the uncertainty in η can modeled using a prior distribution which is parameterized by the VCL for a rhythm, ψ_t . We define a prior distribution on ψ_t which is conditioned on the information in the literature the VCL for various types of conditions, D .

$$\psi_t \sim p_{\psi_t}(\psi_t \mid D) \quad (4.6)$$

In this context, D is a value of the parameters which determine the frequency of a particular type of heart condition and the mean and standard deviation of the ranges of VCL for a condition. For example, the ranges of the VCL for Ventricular Tachycardia (VT) in a clinical trial conducted prior to RIGHT was $257.3 \pm 41.2[ms]$. We define the complete prior distribution for EGM as:

$$\begin{aligned} X_j &\sim p_X(x_j \mid \eta) \\ \eta &\sim p_\eta(\eta \mid D) \end{aligned} \quad (4.7)$$

where, for notation, we have defined η to include ψ_t and the prior as $p_\eta(\eta \mid D)$.

To account for the uncertainty in the parameters we integrate:

$$p_X(x | D) = \int p_X(x | \eta) p_\eta(\eta | D) d\eta \quad (4.8)$$

The generated cohorts are then applied to the device which we can model as independent Bernoulli random variable Y_j with parameters θ_d ($d \in \{V2, MDT\}$), and X_j . Y_j takes the following values:

$$Y_j = \begin{cases} 1, & \text{output is inappropriate therapy} \\ 0, & \text{output is not inappropriate therapy} \end{cases} \quad (4.9)$$

Estimate of inappropriate therapy rate

From the responses of the device model, we can define the likelihood function for a device:

$$L(\theta_d | y_o, x_o) = \prod_{j=1}^{N_o} p_Y(y_j | \theta_d, x_j) p_{X_j}(x_j | D) \quad (4.10)$$

where x_o, y_o is a shorthand notation for the N_o EGMs and device outputs. Similar to (4.8), the variability in the generated cohort can be integrated out:

$$L(\theta_d | y) = \int \prod_{j=1}^{N_o} p_Y(y_j | \theta_d, x_j) p_X(x_j | D) dx_j \quad (4.11)$$

There is no a closed-form solution for this expression, however we use Monte-Carlo methods to estimate θ (Gelman et al., 2014).

4.5.3. Effect of incorporating prior data on VCL distribution

Fig. 32 depicts the VCL of SupraVentricular Tachycardia (SVT) and VT in both the real patient training data set and within the generated virtual cohort. Information available in (Berger et al., 2006) was incorporated according to the (4.8). The mean and standard deviation for one instance of the generated virtual cohort was $485.9 \pm 118[ms]$ for SVT and

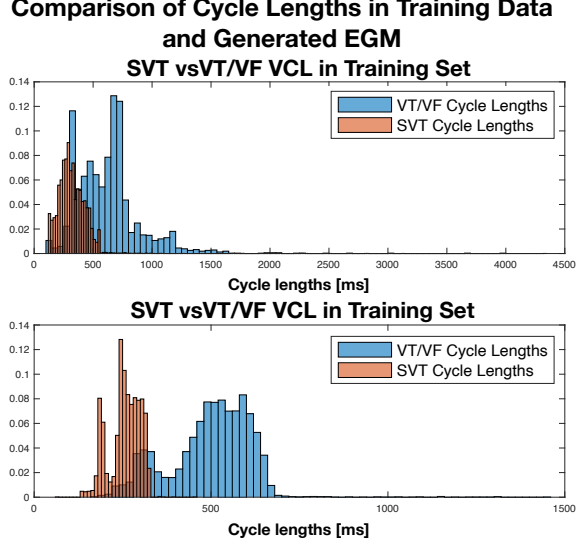


Figure 32: Distribution of VCL for generated virtual cohort. Historical information is accurately reflected to bias the generated VCL for both SVT and VT episodes.

$258.0 \pm 43[ms]$ for VT. This is in comparison to the mean and $625.55 \pm 281.27[ms]$ for SVT and $322.10 \pm 137.11[ms]$ for VT within the real patient training data set. This confirms that the virtual cohort generated reflected information that would have been available at the design of RIGHT.

Estimates of inappropriate therapy rate. We generated a virtual cohort with prior information incorporated, with 11,400 total EGMs (19 different conditions, 600 heart model instances simulated for 50 seconds) and obtained the corresponding responses as before. The generation procedure was repeated for a total of 100 iterations to obtain an overall estimate of the inappropriate therapy rate.

By only assuming a uniform distribution on the occurrence of arrhythmia within a cohort, analogous to what would be assumed at the design stage of RIGHT, we obtained results that the Vitality 2 (V2) discrimination algorithm had a higher rate of inappropriate therapy than the Medtronic (MDT) discrimination algorithm (33.22% vs 15.62% with p-value < 0.001). Without the prior information, as in the previous section, the results are statistically significant, but the difference in effect size is not as pronounced (9.99% vs 3.88% with p-

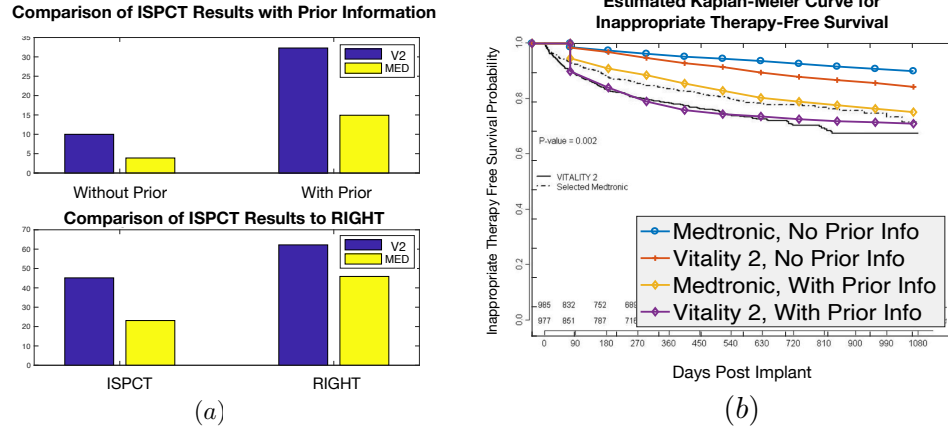


Figure 33: Results of Bayesian hierarchical model for virtual cohort generation: (a) Comparison of Inappropriate Therapy Rates (top) Comparison of ISPCT inappropriate therapy rate estimate with and without prior information. Uniform distribution of arrhythmia type assumed (bottom) Comparison of ISPCT estimate of inappropriate therapy rate to results from RIGHT. (b) Estimated RIGHT Survival Curves.

value < 0.001). This would lead to greater uncertainty from the results. This comparison is shown in Fig. 33 (a, top).

In order to further validate the ISPCT outcomes, the distribution heart conditions from the results of RIGHT (Gold, Ahmad, et al., 2012) was utilized to estimate the inappropriate therapy rate, retrospectively. With this additional information, effect sizes of 45.6% vs 23.11% for V2 vs. MDT were observed. Thus, further reducing the gap between the results of the ISPCT and the results of RIGHT which were 62.2% vs 45.9% (Fig. 33 (a, bottom)). This is an approximately 42% increase in accuracy of the predicted inappropriate therapy rate.

4.5.4. Mapping the ISPCT results to RIGHT

Assuming that the distribution of heart conditions remain constant, the time-to-first inappropriate therapy, the original endpoint of RIGHT, can be estimated using a geometric distribution, whose parameter is the inappropriate therapy rate, θ . We utilize the study in (Fontenla, 2016), to obtain information about the distribution of heart conditions and the rate of occurrence within a population, for example, for SVT: From the cohort of 1514 enrolled patients, 428 had 2596 non-ventricular SVT episodes, assuming a constant rate of

occurrence, the average SVT per patient is:

$$\text{SVT per patient} = 2596/428 = 6.0654$$

$$\text{Interval of SVT events} = 3 \text{ years} * 365 / \text{SVT per patient} = 180.5 \text{ [days]}$$

Sampling from the geometric distribution with parameter set to θ for each device, we plot these according to time using the values above and obtain the survival curves in Fig. 33 (b). The shape of the estimated curve is comparable to the actual survival curve of RIGHT. The estimated curve without the historical information exhibits a large difference in shape.

After fitting a Cox proportional hazard curve to the estimates, we obtained a hazard ratio of 1.29677 ($P < 0.001$) This was in comparison to a hazard ratio of 1.63 ($P < 0.001$) obtained in RIGHT. The difference in the hazard ratio obtained with the prior information incorporated is significantly less compared to the discrepancy in hazard ratio without the prior information, which was 2.196 ($P < 0.001$).

4.6. Chapter Conclusion

In this chapter, we explored the necessity and the advantage of using physiological simulation for the evaluation of a medical device. As our main application, we presented the requirements for a physiological EGM simulator for the ICD. This simulator was used to create a virtual cohort of EGMs which was then applied to a device model of the ICD discrimination algorithm in order to create a virtual cohort of endpoints. This cohort was used to evaluate the ICD discrimination algorithm of two major ICD manufacturers and the results in terms of relative inappropriate therapy rate was shown to be aligned with the result of RIGHT. In addition, the usage of physiological simulation and virtual cohort generation allows for condition-level analysis that helps determine which algorithm is more susceptible to errors when certain conditions are present in the signal. Moreover, effects of device settings on the performance of the device can also be measured using the virtual cohort. These results were also in alignment with outcomes in other ICD clinical trials.

The virtual cohort generation process was further developed using a hierarchical model which allowed for the incorporation of historical information in a systematic manner. The results demonstrated the improvement in accuracy in terms of the absolute estimates for inappropriate therapy rate for the ICD.

In addition, the results of the ISPCT could be mapped back to the original endpoint of the RIGHT trial, albeit some assumptions. In this case as well, incorporating historical information allowed for more accurate estimates of the time to inappropriate therapy.

Through the results in this chapter, we addressed the challenge of generating a virtual cohort and demonstrated that it can be used to predict the outcomes of a clinical trial.

However, the challenge remains about how to utilize virtual cohorts as an additional source of information within the context of a clinical trial and combine it with real data from a clinical trial to improve the power. Moreover, how to account for the uncertainty due to the usage of simulated data is unclear. We address these challenges in the next chapter.

CHAPTER 5 : Computer-aided Clinical Trials and Uncertainty Quantification

5.1. Overview

In this chapter, we address the issues of evaluating a medical device with a virtual cohort and considering the uncertainty in the outcomes due to the use of simulated data. This work has been adapted from “Computer Aided Clinical Trials for Implantable Cardiac Devices” in the Proceeding of the Conference of the Engineering in Medicine and Biology Society 2018 and “Robustness evaluation of computer-aided clinical trials for medical devices” in the Proceedings of the International Conference on Cyber-Physical Systems. These papers were joint work in collaboration with Yash Vardhan Pant, Bo Zhang, James Weimer, Houssam Abbas, Zhihao Jiang, Jackson Liang, Sanjay Dixit, and Rahul Mangharam.

5.2. Introduction

From in-silico evaluation to computer-aided clinical trials. Up until now, we have demonstrated how a virtual cohort that is generated using the simulation of a physiological model can be used to evaluate a medical device. By incorporating hierarchical modeling techniques, we were able to estimate various performance metrics with some accuracy. However, utilizing this additional source of information in the context of a clinical trial, similar to how historical controlled trials incorporate prior results, requires additional considerations.

Computer models of physiology and simulation can potentially be utilized in clinical trials by considering them as an alternative source of prior information (Food, Administration, et al., 2011; G. Campbell, 2017). As we have seen in the prior chapter, the endpoints are simulated using a physiological model and subsequently applied to either a device model or the physical device itself. However, physics-based models are difficult to develop and even when they are available, incur large variability in the simulated endpoints due to the uncertainty in the parameters (Cobelli, Renard, and Kovatchev, 2011). Moreover, these simulations are based on assumptions and can cause large errors in the evaluation of a device when the assumptions differ from reality.

Therefore, a major obstacle in the usage of simulated endpoints as a source of prior information in clinical trials is the lack of a method to quantify the uncertainty caused by assumptions in a simulation and the effect on clinical trials outcomes.

Expanding on the definition provided in (H. Abbas, Zhihao Jiang, et al., 2016), we define a Computer-Aided Clinical Trial (CACT) as a process in which high-level models of human physiology are used to evaluate device performance through the utilization of simulated endpoints called *virtual cohorts* obtained from the simulation of physiological models in the planning, conduct and analysis of a clinical trial. This definition reflects the potential for incorporating simulation results in all stages of a medical device trial.

In the remainder of this chapter, we develop the framework of CACTs by resolving the following issues:

1. How do we model the uncertainty and assumptions of a computer-aided clinical trial?
2. How can we utilize virtual cohorts as an additional source of information and also combine them with real data which would be observed in a clinical trial?
3. Finally, we address the issue of quantifying the uncertainty in the outcome of a CACT.

First, we develop a Bayesian statistical framework for modeling a CACT and its outcome. The framework incorporates results from simulated endpoints through the *pre-clinical prior* and explicitly models simulation assumptions, which will be used in order to quantify the uncertainty of an CACT's outcome. Within the context of our main application of evaluating the ICD discrimination algorithm, we provide results demonstrating the utility of the framework and how a CACT can improve the overall power of a medical device trial. Finally, we define a measure of uncertainty, called δ -robustness, which allows for us to quantify the uncertainty in a CACT's outcome. We demonstrate how the proposed δ -robustness metric can be estimated and utilized in the analysis of CACT outcomes.

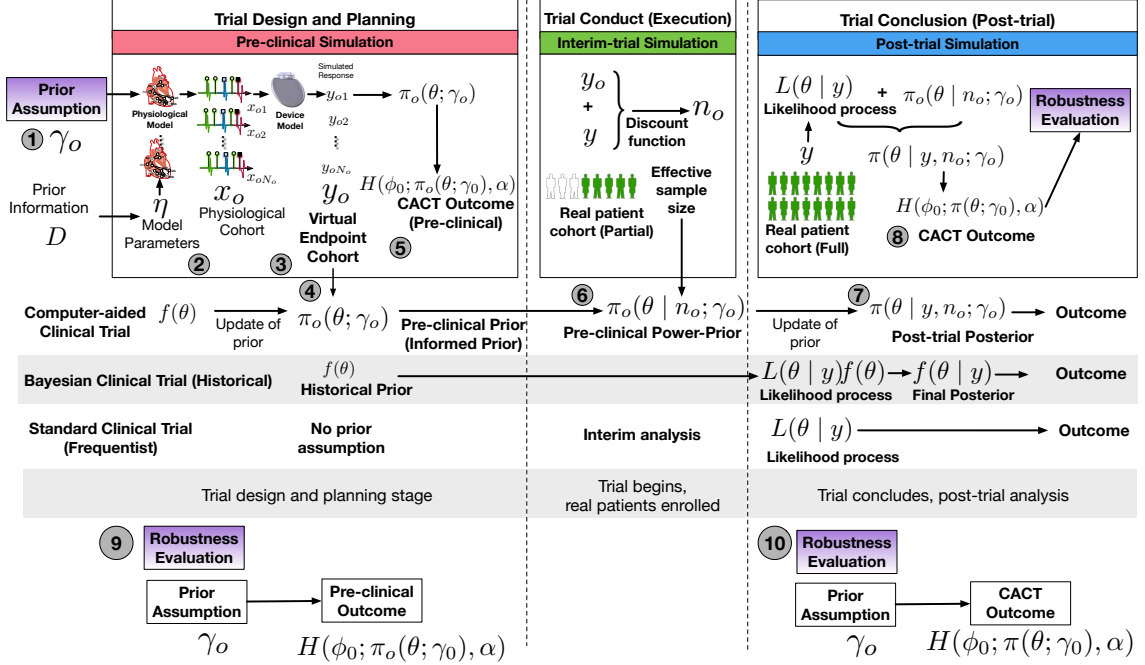


Figure 34: Overview of a Computer-Aided Clinical Trial (CACT) for medical devices and robustness evaluation. A CACT is divided into three phases: pre-clinical simulation (trial design and planning), interim-trial simulation (trial conduct), and post-trial simulation (trial conclusion)

5.3. Modeling Uncertainty and Assumptions in a CACT

5.3.1. Overview of CACTs for medical devices

We first recap the key concepts in a clinical trial for medical devices presented in Chapter 3 and expand upon them. A clinical trial evaluates a medical device through an inquiry in the form of a hypothesis test about a parameter of interest, θ , with a set of observations called *endpoints*, $y = \{y_1, y_2, \dots, y_N\}$ from a patient cohort of sample size N . Examples of endpoints include continuous endpoints, such as the amount of blood pressure reduction, and discrete endpoints, such as the number of inappropriate therapies. The typical hypothesis test compares the likelihood of a null hypothesis, H_0 , with one or more alternative hypotheses, H_1, H_2, \dots , etc. The *outcome* of the trial is either the rejection of the null hypothesis in favor of an alternative hypothesis or the failure to reject the null hypothesis.

As shown in Fig. 34, the clinical trial can be divided into three stages: trial planning, trial conduction, and trial conclusion. Two general approaches to clinical trial evaluation exist.

The standard, frequentist approach to a clinical trial considers only the observed endpoints of the trial cohort to answer the primary question. In a Bayesian clinical trial, a prior distribution (e.g. non-informative uniform prior or historical control-based prior) is placed on the parameter of interest during the planning stage and then updated to form a posterior distribution using observations from the trial. The trial outcome is determined with this posterior distribution.

Similar to the standard clinical trial, a CACT is divided into three stages paralleling the stages of a clinical trial: pre-clinical simulation, interim-trial simulation, and post-trial simulation. During pre-clinical simulation, a physiological model is used to generate a *virtual cohort of physiological signals* (Fig. 34 ②) and is applied to a *device model* in order to generate a cohort of size N_o of simulated endpoints, $y_o = \{y_{o1}, y_{o2}, \dots, y_{oN_o}\}$, called the *virtual endpoint cohort* or *virtual cohort* (Fig. 34 ③).

These endpoints are used to update a non-informative prior distribution to form an informed prior, which we call the *pre-clinical simulation prior distribution* or *pre-clinical prior* (Fig. 34 ④). At the pre-clinical simulation stage, the pre-clinical simulation prior is used to predict the *pre-clinical CACT outcome* of the trial (Fig. 34 ⑤).

The pre-clinical prior is updated when data from real patients become available during the trial conduction and at the conclusion. The pre-clinical prior is weighted according to the similarity between the virtual cohort and the real patient cohort to form the *pre-clinical power prior* (Fig. 34 ⑥).

During post-trial simulation, the *post-trial simulation posterior distribution* is obtained by updating the pre-clinical power prior (Fig. 34 ⑦). Finally, the overall *post-trial simulation CACT outcome* is computed (Fig. 34 ⑧).

In the case of standard clinical trials, assumptions are made about the target population in order to compute parameters of the trial, such as the sample size. Similarly, in a CACT, simulation assumptions, such as the composition of conditions in a virtual cohort, are

modeled within the framework as the CACT prior distribution (Fig. 34 ①). Sec. 5.5 describes the concept of *robustness of CACT outcomes* and how to evaluate the effect of the CACT prior on the CACT outcome (Fig. 34 ⑨,⑩).

In the next section, we formally define the concepts presented in this section within a Bayesian statistical framework.

5.3.2. Bayesian modeling framework for CACTs

In this section, we have expanded the framework proposed in (Haddad, Himes, Thompson, et al., 2017) in order to reflect the different stages of a CACT and explicitly model sources of uncertainty.

Given a set of simulated endpoints, $y_o = \{y_{o1}, y_{o2}, \dots, y_{oN_o}\}$, where $Y_o \sim p_{Y_o}(y_o \mid \theta_o; \gamma_o)$ and a set of real patient cohort endpoints $y = \{y_1, y_2, \dots, y_N\}$, where $Y \sim p_Y(y \mid \theta)$, the conclusions drawn from the CACT regarding θ rely on the following assumption about the exchangeability of the two sets of endpoints:

Assumption 1 (Exchangeability of simulation endpoints). The device model and the physiological model accurately capture the variability of outcomes and generate simulated endpoints, y_o , such that,

$$\theta = \theta_o \tag{5.1a}$$

$$\text{and } p_{Y_o}(y_o \mid \theta; \gamma_o) = p_Y(y \mid \theta) \tag{5.1b}$$

This implies that the information about the parameter of interest obtained through the virtual cohort is the same as that of an actual patient cohort. Moreover, (5.1b) implies that the distribution of the endpoints is equivalent. (5.1a) and (5.1b) must be verified after the start of the trial when endpoints from real patients, y , becomes available. Here, γ_o is a parameter related to the simulation assumption which will be defined in the next section.

CACT pre-clinical simulation We first define the virtual cohort and the relation to the simulation assumptions in the pre-clinical stage and how the assumptions influence the pre-clinical trial outcome. We assume the existence of a device model and an adequate physiological model which generates a cohort of size N_o of physiological signals, $x_o = \{x_{o1}, x_{o2}, \dots, x_{oN_o}, \}$ (Fig. 34 ②) and define the *virtual physiological cohort*:

Definition 1 (Virtual physiological cohort). The virtual physiological cohort or physiological cohort, x_o , is a set of size N_o consisting of I.I.D. instances of the multivariate random variable, X_{oi} , where,

$$X_{oi} \sim p_{X_{oi}}(x_{oi} \mid \psi) \quad (5.2)$$

$p_{X_{oi}}$ is the distribution of X_{oi} indexed by the parameter ψ and is called the distribution of physiological signals conditioned on ψ . Therefore, for the set $X_o = \{X_{o1}, X_{o2}, \dots, X_{oN_o}\}$,

$$p_{X_o}(x_o \mid \psi) = \prod_{i=1}^{N_o} p_{X_{oi}}(x_{oi} \mid \psi) \quad (5.3)$$

The physiological cohort represents a set of signals that are used by the device, such as the cardiac electrograms for a pacemaker or the blood glucose level in an insulin pump. Typically, there is no closed form for p_{X_o} , but samples can be obtained from the distribution by simulating the physiological model. The cohort generation process is equivalent to obtaining samples from the distribution of X_o for which ψ is a parameter setting in that process. For example, if the physiological model generates k types of cardiac rhythms, $\psi = (\psi_1, \psi_2, \dots, \psi_k)$ can be the number of each rhythm type in the physiological cohort. The proportion of each rhythm type in a physiological cohort is an example of a simulation assumption. This assumption is modeled within a Bayesian framework as a prior distribution on the parameter ψ , which we define as the *CACT prior distribution*:

Definition 2 (CACT prior distribution). The parameter Ψ of the virtual cohort is a random variable such that,

$$\Psi \sim \pi(\psi \mid \gamma_o) := \pi_{\gamma_o}(\psi) \quad (5.4)$$

where, γ_o is the set of parameters for the prior distribution. $\pi_{\gamma_o}(\psi)$ is defined as the CACT prior distribution.

$\pi_{\gamma_o}(\psi)$ encodes information of simulation assumptions by specifying the form of the function and parameters γ_o . This information may be available from reports of previous clinical trials, otherwise a conservative base prior may be selected, such as the uniform distribution.

With this prior, we define the marginal distribution of physiological cohorts, $p_{X_o}(x_o; \gamma_o)$ as:

$$p_{X_o}(x_o; \gamma_o) = \int p_{X_o}(x_o \mid \psi) \pi_{\gamma_o}(\psi) dF(\Psi) \quad (5.5)$$

This can be considered the predictive prior distribution in Bayesian analysis. Note, the integration is over the support of the random variable Ψ . For simplicity of notation, we will suppress the upper and lower limits throughout the paper.

Intuitively, the marginal distribution obtains a weighted average of x_o , where the weight is determined by the CACT prior distribution, $\pi_{\gamma_o}(\psi)$. The value of the parameter, γ_o , influences the marginal distribution of cohorts.

A closed form of $p_{X_o}(x_o; \gamma_o)$ is typically unavailable, but we assume that obtaining instances from the distribution of physiological cohorts, x_o , is possible. Therefore, marginalization is implemented using sampling schemes, such as Monte Carlo methods.

The virtual cohort, y_o , (Fig. 34 ③) is obtained by applying each of the signals in the physiological cohort to the device model:

Definition 3 (Virtual endpoint cohort). Given a physiological cohort x_o , the virtual endpoint cohort or virtual cohort, y_o , is a multivariate random variable Y_o of dimension N_o , such that,

$$Y_o \sim p_{Y_o}(y_o \mid x_o, \theta_o; \gamma_o) \quad (5.6)$$

where p_{Y_o} is the distribution of Y_o indexed by θ_o and γ_o is the parameter value for the CACT prior distribution.

Here, θ_o is value of the parameter of interest that is obtained with the virtual cohort. Note, the endpoints comprising the virtual cohort are simulations of the outcomes in a patient cohort, such as the number of inappropriate therapy for a device, and not the physiological signals. As before, the marginal distribution $p_{Y_o}(y_o \mid \theta_o; \gamma_o)$ is defined as:

$$p_{Y_o}(y_o \mid \theta_o; \gamma_o) = \int p_{y_o}(y_o \mid x_o, \theta_o; \gamma_o) dF(X_o) \quad (5.7)$$

From the virtual cohort, we wish to obtain estimates of θ . Assuming (5.1a) and (5.1b), for a single virtual cohort y_o , we define the likelihood function with respect to θ as $L(\theta \mid y_o; \gamma_o) = p_{Y_o}(y_o \mid \theta; \gamma_o)$.

Following a typical Bayesian procedure, we place a minimally-informative prior on θ , $\pi_o(\theta)$, such as the uniform prior, and define the posterior distribution of θ , $\pi_o(\theta \mid y_o)$ for a single virtual cohort y_o such that:

$$\pi_o(\theta \mid y_o; \gamma_o) \propto L(\theta \mid y_o; \gamma_o) \pi_o(\theta) \quad (5.8)$$

In cases when $\pi_o(\theta)$ and p_{Y_o} are conjugates, then we have the closed-form of the posterior:

$$\pi_o(\theta \mid y_o; \gamma_o) = p_{Y_o}(y_o \mid \theta; \gamma_o) \pi_o(\theta) \quad (5.9)$$

As before, we marginalize out the uncertainty from y_o to obtain the *pre-clinical prior distribution of θ* (Fig. 34 ④):

Definition 4 (Pre-clinical simulation prior distribution). For the parameter of interest θ and (5.4) with parameter set to γ_o , the pre-clinical simulation prior distribution or pre-clinical prior distribution for θ is defined as,

$$\pi_o(\theta; \gamma_o) = \int \pi_o(\theta \mid y_o; \gamma_o) dy_o \quad (5.10)$$

Similar to before, this can be thought of as a weighted average of θ , where the weights are determined according to the distribution of y_o . Note, (5.10) is an informed prior and is different from the CACT prior.

Finally, the pre-clinical prior distribution is used to determine the pre-clinical simulation outcome. Consider evaluating the assertion $\theta \in \phi_0$, i.e., the parameter of interest falls in the region ϕ_0 . For instance, when evaluating two medical devices, this could mean: the difference of inappropriate therapy rate of Device A and Device B is non-negative (A is worse than B), i.e., $\theta \geq 0$. Here $\phi_0 = [0, \infty)$.

We define a function which encodes the results of evaluating the assertion $\theta \in \phi_0$ based on the posterior distribution of θ as the outcome of pre-clinical simulation (Fig. 34 ⑤):

Definition 5 (Pre-clinical simulation CACT outcome).

$$H(\phi_0; \pi_0(\theta; \gamma_0), \alpha) = \begin{cases} 1 & \text{if } P(\pi_0(\theta; \gamma_0) \in \phi_0) \geq 1 - \alpha \\ 0 & \text{o.w.} \end{cases} \quad (5.11)$$

Here, an assertion of $\theta \in \phi_0$ means that at least $(1 - \alpha)\%$ of the posterior distribution of θ is supported on ϕ_0 . The relation to the assumption is denoted by γ_0 . For the previous

example, with $\alpha = 0.05$, $H(\phi_0; \pi_0(\theta; \gamma_0), \alpha) = 1$ means that the probability that Device A has a higher inappropriate therapy rate than Device B is at least 95%.

In addition to the outcome itself, (5.10) can be used to estimate various statistics, such as the mean and variance of θ which is used to derive parameters relevant to the design of the clinical trial, such as the desired sample size N .

5.4. Combining Virtual Cohorts and Real Patient Data

Next, we describe how to combine the results from pre-clinical simulation with endpoints from real patients available during trial conduct and at the conclusion of the trial. This pertains to the CACT interim simulation and post-trial simulation stages of Fig. 34. In order to combine the virtual cohort y_o with the real cohort y , the power-prior framework Joseph G Ibrahim et al., 2015a is applied to virtual cohorts to define the pre-clinical simulation power-prior distribution for a single cohort y_o :

Definition 6 (Pre-clinical simulation power-prior distribution for y_o). For a single virtual y_o of size N_o and real patient outcomes y , the pre-clinical simulation power-prior distribution for y_o is defined as,

$$\pi_o(\theta \mid y_o, n_o; \gamma_o) \propto L(\theta \mid y_o; \gamma_o)^{\frac{n_o}{N_o}} \pi_o(\theta) \quad (5.12)$$

where n_o , the effective virtual cohort sample size is determined through a discount function which evaluates the similarity of y and y_o .

5.4.1. Discount function for effective virtual cohort sample size

An example of a discount function and a method for determining the parameters of the function is proposed in Haddad, Himes, and M. Campbell, 2014. Fig. 35, illustrates how the discount function monitors the similarity between the virtual cohort and the real cohort by outputting a *p-value*. In the example, the real cohort follows a Binomial distribution with proportion parameter θ ranging from 0.45 to 0.6. When the virtual cohort is sampled

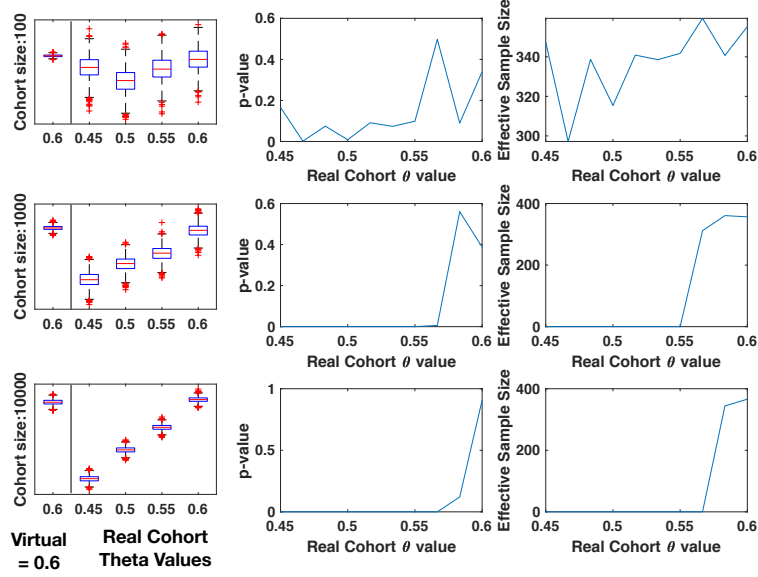


Figure 35: Monitoring of CACT prior using discount function. As the virtual cohort becomes similar to the real cohort, the p -value increases and the corresponding effective sample size n_o increases. With a larger real cohort, the effective sample size only increases with a strong degree of similarity.

from a Binomial distribution $Bin(N, p = 0.6)$ the p -value output from the increases as the real θ value approaches 0.6 and the effective sample size, n_o increases accordingly. Also, as observed in Fig. 35, as the real cohort size is increased from 100 to 10000 the influence of the virtual cohort is only apparent when the two cohorts are more similar.

We apply the procedure in a case study described in Sec. 5.6. A discussion about choosing the appropriate parameters for the discount function is provided in Sec. 5.7.1.

As in (5.7), the uncertainty in the virtual cohort y_o is marginalized out (Fig. 34 ⑥):

Definition 7 (Pre-clinical simulation power-prior distribution).

$$\pi_o(\theta \mid n_o; \gamma_o) = \int L(\theta \mid y_o; \gamma_o)^{\frac{n_o}{N_o}} \pi_o(\theta) dF(y_o) \quad (5.13)$$

(5.13) can be interpreted as a weighted average over the different virtual cohorts y_o , where the weights are determined according to the distribution of y_o .

For the real patient outcomes y , we define the likelihood $L(\theta|y) = p_Y(y|\theta)$, according to assumption (5.1b). $L(\theta|y)$ and (5.13) are combined to form the post-trial posterior distribution (Fig. 34 ⑦):

Definition 8 (Post-trial simulation posterior distribution).

$$\pi(\theta \mid y, n_o; \gamma_o) \propto L(\theta \mid y) \pi_o(\theta \mid n_o; \gamma_o) \quad (5.14)$$

If $p_Y(y \mid \theta)$ and $\pi_o(\theta; \gamma_o)$ are conjugates then,

$$\pi(\theta \mid y, n_o; \gamma_o) = p_Y(y \mid \theta) \pi_o(\theta \mid n_o; \gamma_o) \quad (5.15)$$

The effective sample size, n_o , can be interpreted as discounting the influence of the virtual cohort y_o on the overall posterior and the outcome if y and y_o differ greatly.

Finally, based on the post-trial posterior distribution $\pi(\theta \mid y, n_o; \gamma_o)$, the post-trial outcome of the CACT is defined (Fig. 34 ⑧):

Definition 9 (Post-trial simulation CACT outcome).

$$H(\phi_0; \pi(\theta; \gamma_0), \alpha) = \begin{cases} 1 & \text{if } P(\pi(\theta; \gamma_0) \in \phi_0) \geq 1 - \alpha. \\ 0 & \text{o.w.} \end{cases} \quad (5.16)$$

In Sec. 5.6, we apply the framework to the Rhythm ID Goes Head-to-head Trial (RIGHT) in order to demonstrate the utility of the framework.

5.5. The δ -Robustness of CACT Outcomes

5.5.1. *Measuring uncertainty with respect to sensitivity to assumptions*

With the framework defined above, we can begin to address how uncertainty in the outcome can be quantified. Many approaches to uncertainty quantification exist in the literature. Of those, Bayesian sensitivity analysis aims to analyze and quantify the affect of prior distributions on the outcomes of analysis. For example, (Roos et al., 2015) presents a method for quantifying the sensitivity of the posterior distribution in a hierarchical model to the prior distribution using the Hellinger distance. A limitation of these approaches is that they are difficult to interpret, especially in their relevance to a clinical trials. Therefore, in this section, we propose a new measure for quantifying the uncertainty in a clinical trial outcome.

Ideally, we would like a measure that would allow us to determine that a CACT outcome is ‘robust’ to ‘large’ perturbations in the CACT prior distribution. This would lead to an increased confidence in the CACT outcomes. We formalize the the concepts of *large* and *robust* by defining ϵ -perturbations and δ -robustness.

The sensitivity of the CACT outcome to a base prior distribution can be determined by ‘perturbing’ the prior distribution by an ϵ amount and evaluating the effect on the outcome.

For parametric distributions, π_γ , this is accomplished by perturbing the parameter γ :

Definition 10 (ϵ -perturbation of π_{γ_o}). Let π_{γ_o} be the base distribution with base parameter γ_o and $\pi_{\gamma(\epsilon)}$ be the perturbed distribution, with parameter set to $\gamma(\epsilon)$. Given a distance measure between two probability distributions $D(P, Q)$, an ϵ -perturbation of π_{γ_o} is a set of parameters $\gamma(\epsilon)$ such that,

$$D(\pi_{\gamma_o}, \pi_{\gamma(\epsilon)}) = \epsilon \tag{5.17}$$

From this, we define the robustness of a CACT outcome:

Definition 11 (δ -robustness of a CACT outcome). Given a base prior distribution $\pi_{\gamma_o}(\cdot)$ and the CACT outcome for a parameter of interest θ , $H(\cdot, \pi(\theta; \gamma_o))$, the CACT outcome has a robustness of δ if the following condition holds:

$$\begin{aligned} \forall \gamma \in \{\gamma(\epsilon) : \epsilon \in [0, \delta]\}, \\ H(\cdot, \pi(\theta; \gamma_o)) = H(\cdot, \pi(\theta; \gamma)) \end{aligned} \tag{5.18}$$

We now formally define the estimation problem as follows:

Problem (Estimation of the δ -robustness value of a CACT outcome). Given a base prior distribution $\pi_{\gamma_o}(\cdot)$ and function for the CACT outcome $H(\cdot, \pi)$, determine,

$$\delta = \operatorname{argmin}_{\epsilon} H(\cdot, \pi(\theta; \gamma_o)) \neq H(\cdot, \pi(\theta; \gamma(\epsilon))) \tag{5.19}$$

Here, the range of ϵ depends on the distance measure used. The resulting δ is defined as the δ -robustness value.

Intuitively, starting from a base prior distribution, the δ -robustness value is the minimal perturbation of magnitude δ required to detect a change in the CACT outcome. Fig. 36 depicts how the CACT outcome with a base prior distribution is initially 1. As the magnitude of perturbation (ϵ) increases, the outcome remains constant until δ_o is reached. At δ_o , the outcome changes from the initial positive outcome to 0. In this example, the δ -robustness for the CACT outcome is δ_o . Our procedure for δ -robustness estimation is presented in the next section.

5.5.2. Procedure for estimation of δ -robustness

Depending on the choice of distance measure, as well as the form of the base prior distribution, the δ -robustness value for a CACT outcome will vary. In this work, we use the

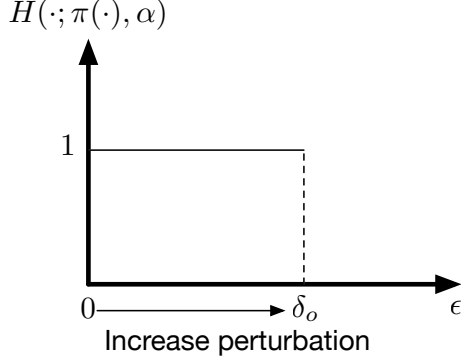


Figure 36: δ -robustness of outcome $H(\phi; \pi(\cdot), \alpha)$. Perturbations ϵ of π_{γ_o} are increased until a magnitude of δ_o , at which the outcome differs from the initial outcome.

Hellinger distance (Le Cam, 1986) between two discrete distributions:

Definition 12 (Hellinger distance between two discrete distributions). For two discrete distributions $\pi_{\gamma_o} = (\gamma_{o1} \dots \gamma_{ok})$ and $\pi_{\gamma} = (\gamma_1 \dots \gamma_k)$, the Hellinger distance is defined as,

$$D_H(\pi_{\gamma}, \pi_{\gamma_o}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{\gamma_i} - \sqrt{\gamma_{oi}})^2} \quad (5.20)$$

In this work, we estimate the δ -robustness with respect to the base distribution (e.g. CACT prior) using the Hellinger distance. One advantage to using the Hellinger distance include is that it is a symmetric measure of distance, unlike the Kullback-Liebler divergence. Other advantages are described in (Roos et al., 2015). The functional form of the distance also enables convenient numerical computation of distributions which are at ϵ distance (see Appendix A.3 for details).

Procedure

1. Evaluate the CACT outcome with the base distribution π_{γ_o}
2. Starting from a small ϵ and gradually increasing the perturbation over the range of $\epsilon \in (0, 1]$, for each ϵ , sample multiple perturbed distributions $\pi_{\gamma(\epsilon)}$ and evaluate the

CACT outcome.

3. Continue until the CACT outcome differs from the initial outcome.

There exist situations where the assumption of the CACT does not affect the outcome, trivially. In this case, the robustness would be infinite, which is ambiguous in meaning. For this reason, we define the weighted δ -robustness:

Definition 13 (Weighted δ -robustness). Let $W(\delta)$ be a weight with real value. Then, the weighted δ -robustness takes the value as defined below,

$$\text{Weighted } \delta\text{-robustness} = \begin{cases} \delta W(\delta), & \text{if } W(\delta) > 0 \\ \text{N.A.}, & \text{otherwise} \end{cases} \quad (5.21)$$

The weight $W(\delta)$ can be defined such if the weight is zero, then that means the assumption has no impact on the outcome.

In (Roos et al., 2015), the δ -local circular sensitivity $S_{\gamma_o}^c(\delta)$ is defined as:

$$S_{\gamma_o}^c(\delta) = \left\{ \frac{D(\pi_\gamma(\theta | y), \pi_{\gamma_o}(\theta | y))}{\delta}, \text{ for } \gamma \in G_{\gamma_o}(\delta) \right\} \quad (5.22)$$

A sensitivity value of 0 indeed indicates that the base prior distribution has no influence over the posterior distribution.

In this paper, we used the weighting function:

$$W(\delta) = E[S_{\gamma_o}^c(\delta)] \quad (5.23)$$

Therefore, the robustness value is weighted by the average of sensitivity at a perturbation of δ . The weighted δ -robustness is zero only if the average weight is zero, which can only

happen if perturbations in the assumption have no effect on the outcome of the CACT. This resolves the ambiguity for cases of infinite robustness.

5.6. Case Study: CACT for RIGHT (CACT-RIGHT)

In this section, we apply the CACT framework to RIGHT (Berger et al., 2006) which we call Computer-Aided Clinical Trial for RIGHT (CACT-RIGHT).

Definition 14 (Parameter of interest for CACT-RIGHT). For each device d , the inappropriate therapy rate for a cohort of size N_o is φ_d , where, $d \in \{0:\text{Vitality II}, 1:\text{Medtronic Devices}\}$. The parameter of interest for CACT-RIGHT is the difference in inappropriate therapy rate and is defined as,

$$\theta = \varphi_0 - \varphi_1 \tag{5.24}$$

Without loss of generality, we will focus on the difference in inappropriate therapy rate.

5.6.1. Pre-clinical simulation for CACT-RIGHT

Physiological cohort generation. The physiological model described in Chapter 4 is used to generate the physiological signals.

Here, we model the physiological cohort x_o of size N_o . The cohort consists of $K = 8$ different types of EGM episodes listed in Table 3. For each type i , the j th signal of type i is I.I.D. such that,

$$X_{oi}^{(j)} \sim p_{X_{oi}}(x_{oi} \mid \eta) \tag{5.25}$$

where η is the vector of parameters of the heart model to generate condition i . The model is then simulated to obtain a time-series of the physiological signal as described in Chapter 4. Fig. 37 shows a Ventricular Fibrillation (VF) signal with a cycle length of 286(ms).

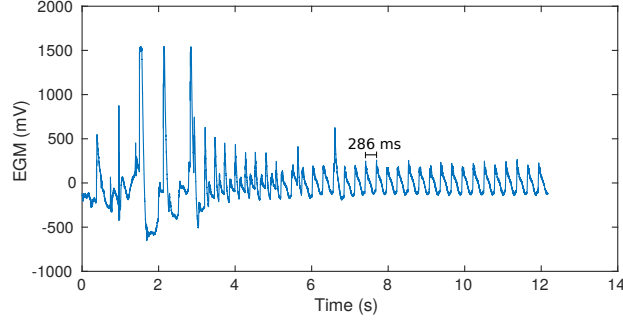


Figure 37: Example of one-channel of the generated physiological signal of type VF with VCL = 286(ms). The signals in the cohort are indexed by the type of condition and the average VCL for a condition.

CACT-RIGHT prior distribution. A single cohort x_o of size N_o is generated as a composition of the different types according to $\psi = (\psi_1, \psi_2, \dots, \psi_K)$, where ψ_i is the number of episodes of type i in the cohort. ψ is a multivariate random variable that is governed by a multinomial distribution defined as,

$$\Psi \sim \pi(\psi \mid \gamma_o) = Mult(N_o, \gamma_o) \quad (5.26)$$

where $\gamma_o = (\gamma_{o1}, \gamma_{o2}, \dots, \gamma_{oK})$ is a discrete probability distribution representing the proportions of each type in the physiological cohort. Setting $\gamma_{oi} = 1/k$, the CACT-RIGHT prior distribution $\pi(\psi \mid \gamma_o)$ assumes a uniform composition of the K types of arrhythmia.

Together with the CACT-RIGHT prior distribution, the distribution of the cohort $f_{X_o}(x_o \mid \psi; \gamma_o)$ becomes:

$$p_{X_o}(x_o \mid \psi; \gamma_o) = \prod_{i=1}^K \prod_{j=1}^{\psi_i} p_{X_{oi}}(x_{oi} \mid \eta) \pi_{\gamma_o}(\psi) \quad (5.27)$$

The marginal distribution of physiological cohorts is (5.5).

Device model and virtual cohort generation. The device model for the two different devices is implemented according to the description of the algorithms in (Boston Scientific Corporation, 2007)(C. D. Swerdlow and others, 2002). Each signal in the physiological

cohort x_o is applied to the device model d to generate a cohort of device outputs, y_{od} . We assume that the device model is a deterministic function for a particular instance of the cohort x_o . This allows for the output y_{od} for a cohort instance to be modeled as a binomial random variable, where y_{od} is the number of inappropriate therapies in a cohort of size N_o :

$$Y_{od} \sim p_{Y_{od}}(y_{od} \mid x_o, \varphi_i; \gamma_o) = \text{Bin}(N_o, \varphi_d) = p_{Y_{od}}(y_{od} \mid \varphi_d; \gamma_o) \quad (5.28)$$

where $d \in \{0, 1\}$ (refer to Def. 14).

We define the likelihood function $L(\varphi_d \mid y_{od}) = p_{Y_{od}}(y_{od} \mid \varphi_i; \gamma_o)$ and assuming a non-informative Beta prior, $\pi_o(\varphi_d) = \text{Beta}(1, 1)$, by conjugacy, the posterior distribution for the inappropriate therapy rate is a Beta distribution such that:

$$\begin{aligned} p_{\varphi_d}(\varphi_d \mid y_{od}; \gamma_o) &= L(\varphi_d \mid y_{od})\pi_o(\varphi_d) \\ &= \text{Beta}(1 + y_{od}, N_o - y_{od} + 1) \end{aligned} \quad (5.29)$$

Sampling over the many cohorts of y_{od} , we obtain the marginal distribution inappropriate therapy rate for device d :

$$p_{\varphi_d}(\varphi_d; \gamma_o) = \int p_{\varphi_d}(\varphi_d \mid y_{od}; \gamma_o) dF(y_{od}) \quad (5.30)$$

(5.30) is a weighted mixture of Beta distributions, for which the weighting is by the distribution of y_{od} .

The conditional distribution of the pre-clinical prior distribution of θ for the CACT-RIGHT can be defined as,

$$\pi_o(\theta \mid \varphi_1, \varphi_2; \gamma_o) \quad (5.31)$$

As before, there is no closed form for this distribution, hence a sampling scheme using

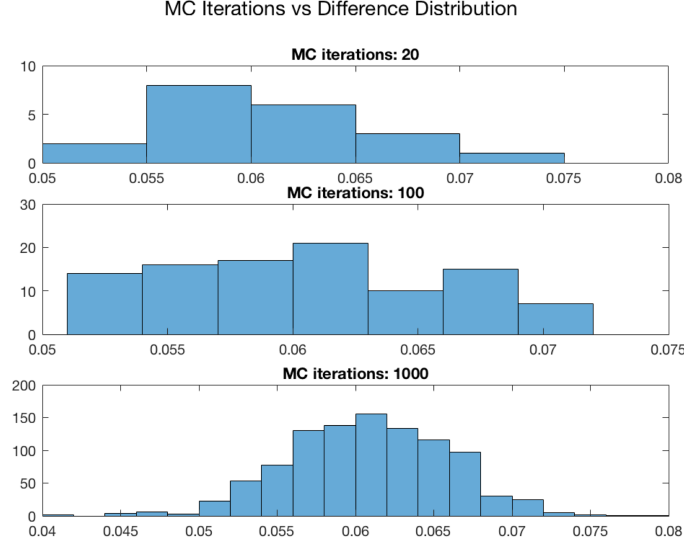


Figure 38: Example of marginalization using Monte Carlo methods to obtain samples from pre-clinical simulation prior.

ordinary Monte-carlo methods is employed. The samples obtained from the posterior distribution form an empirical distribution, as illustrated in Figure 38, from which we can then evaluate the pre-clinical simulation CACT outcome.

CACT-RIGHT pre-clinical simulation outcome. The pre-clinical outcome of CACT-RIGHT is defined with regards to an assertion $\theta \in \phi_0 = \{\theta : \theta < 0\}$, according to the trial. The assertion tests if the inappropriate therapy rate of V2 ICDs is less than MDT ICDs. See (5.24).

δ -Robustness estimation of CACT pre-clinical simulation outcome. The δ -robustness of the pre-clinical simulation outcome with respect to (5.26) is estimated according to the procedure outlined in Sec. 5.5.2.

5.6.2. Pre-clinical simulation results

Fig. 39 shows the results of pre-clinical simulation. During pre-clinical simulation, estimates of the inappropriate therapy rate for Vitality II and Medtronic was $9.99 \pm 0.04\%$ and $3.88 \pm 0.06\%$, respectively. Despite the discrepancy in magnitude, the CACT-RIGHT successfully

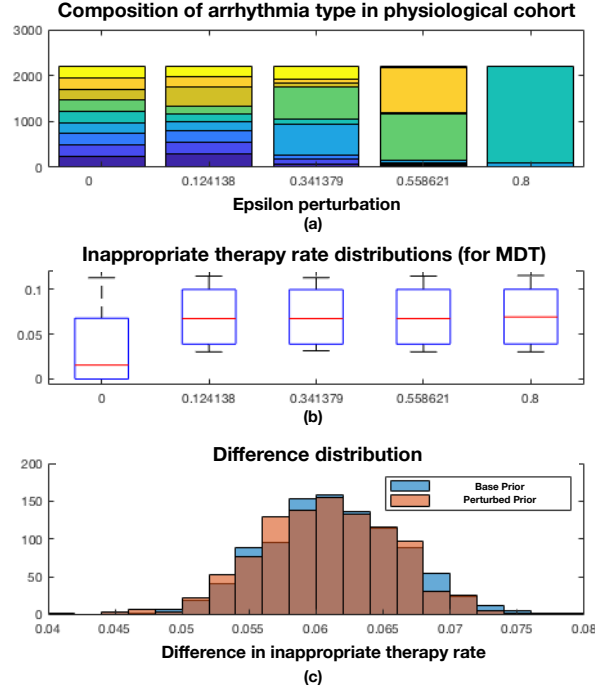


Figure 39: Pre-clinical simulation robustness evaluation. As ϵ -perturbations are increased in magnitude, the composition of the cohort from the base uniform distribution is distorted. (b) As a result of perturbation, the pre-clinical distribution shifts. (c) Pre-clinical prior distribution (blue) and perturbed distribution of $\epsilon = 0.8$ (red). The pre-clinical prior has a large robustness and does not shift. Color in online version.

predicted the effect direction as in RIGHT.

Fig. 39(b) illustrates how perturbing the parameters of the CACT prior affects the distribution of simulated endpoints (inappropriate therapy). From the initial base distribution ($\epsilon = 0$), as the magnitude of the ϵ -perturbation increases, the difference in inappropriate therapy is more pronounced in the same direction. This indicates the CACT-RIGHT outcome is robust with respect to the assumption of a uniform distribution of conditions.

The resulting pre-clinical prior distribution and the distribution after perturbation is shown in Fig. 39(c). With respect to the the CACT-RIGHT prior (5.26), the δ -robustness value of the outcome was 0.778. The maximum robustness in this case is 1, therefore we can conclude that the outcome is relatively large, adding to the confidence of the pre-clinical simulation outcome.

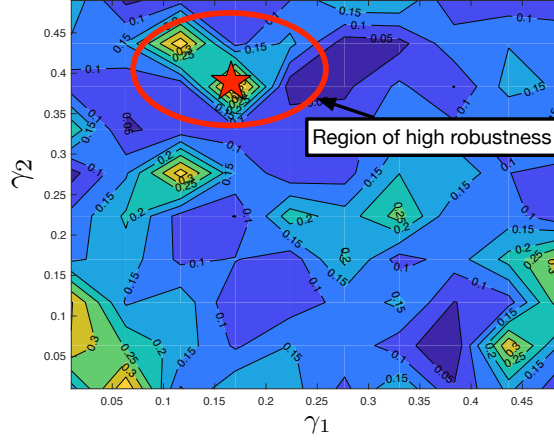


Figure 40: Pre-clinical simulation robustness plane. Many CACT priors are evaluated as the base distribution in order to form the robustness plane. Less emphasis may be placed on recruiting patients corresponding to regions of high robustness (star).

δ -Robustness plane of CACT pre-clinical simulation outcome

For CACT-RIGHT, γ_o represents the average proportion of the different types of arrhythmia in the physiological cohort and the robustness value for other γ_o can be evaluated. For illustration, we evaluated $\gamma_o \in \mathbb{R}^k$, such that γ_{oi} , for $i = 1, 2$ was uniformly distributed in the interval $(0, 0.5)$, and the remaining probability mass was distributed evenly, such that $\gamma_{oj} = (1 - (\gamma_{o1} + \gamma_{o2}))/k$, $\forall j = 3, \dots, k$. The robustness value was computed for each instance of γ_o and plotted to form the *pre-clinical simulation robustness plane*. Fig. 40 shows the weighted δ -robustness plane.

A region on the robustness plane can be interpreted as a subset of the population that exhibits similar proportions of arrhythmias. The red star in the figure indicates a region of higher robustness. Although the robustness value is related to the distribution space, a point with a higher robustness value means there is a relatively larger region in the parameter space of γ_o where the simulation outcome will not change. From a trial planning perspective, this implies that less effort could be put into recruiting patients that correspond to this region, as there is more confidence in the simulation outcome.

Table 3: Summary of RIGHT(Gold, Ahmad, et al., 2012) results. Inappropriate therapy per condition for each of the devices. (Credit: Houssam Abbas)

n episodes (% of total events)		
Adjudicated Rhythm	Vitality II	Medtronic
Ventricular tachycardia	23(1.1)	90(4.6)
Ventricular fibrillation	705(34.9)	994(51.0)
Sinus tachycardia	59(2.9)	220(11.3)
Atrial fibrillation	431(21.3)	101(5.2)
Atrial flutter	66(3.3)	19(1.0)
Atrial tachycardia	20(1.0)	100 (5.1)
Other SVT	178(8.8)	325(16.7)
Sinus rhythm with PVC	18(0.9)	1(0.1)

5.6.3. Interim trial simulation and final results

Simulation of real patient cohort endpoints. Recruitment of the real cohort was emulated by utilizing the results reported at the conclusion of RIGHT (Gold, Ahmad, et al., 2012), summarized in Table 3. By assuming the inappropriate therapy rate will remain constant throughout the trial, we sample from a binomial distribution $Bin(N(t), \varphi_d)$, where $N(t)$ is the number of inappropriate therapies at a time during the trial and φ_d is the final inappropriate therapy rate reported for device d .

The discount function based on the Weibull cumulative distribution $F(p \mid \kappa, \lambda)$ is defined:

$$n0 = n_{max} * F(p \mid \kappa, \lambda) \quad (5.32)$$

where $F(p \mid \kappa, \lambda) = 1 - e^{-(\frac{p}{\lambda})^\kappa}$, p is the ‘p-value,’ and n_{max} is the maximum effective sample size for the virtual cohort (Def. 6). A discussion about selecting λ and κ is in Sec. 5.7.1.

Interim and final results of CACT-RIGHT. Fig. 41 shows the average difference in appropriate therapy rate where the number of episodes increases up to the final number of episodes observed in RIGHT. The figure shows how initially, the mean difference between V2 ICDs and MDT ICDs is less pronounced when the sample size is small, but becomes more pronounced as the cohort size increases. The final difference 9.8 ± 0.2 % is aligned with

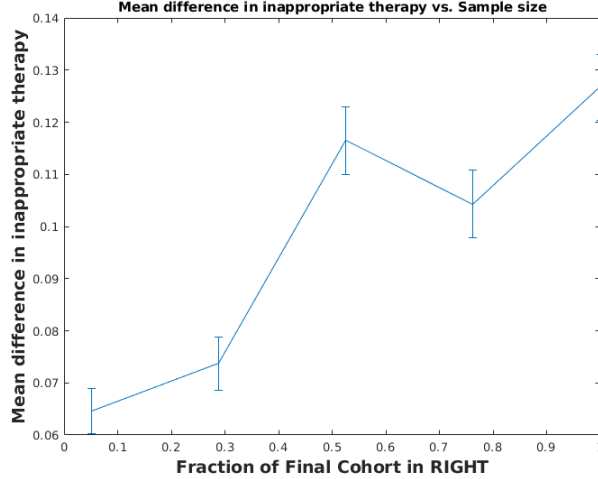


Figure 41: Change in mean difference distribution as more patients enroll. Initially, a small difference in inappropriate therapy rate becomes more pronounced as the cohort size reaches the size of the cohort at the conclusion of RIGHT.

the difference in inappropriate therapy rate 10.1% of RIGHT.

5.6.4. Post-trial simulation robustness analysis

At the conclusion of the trial, information about the CACT prior (i.e. the incidence of the types of arrhythmia) would be available, allowing for a post-trial robustness evaluation to assess how the outcomes may change with a different cohort. We set the parameters for the CACT prior to the values according to Table. 3. The robustness evaluation, Fig. 42(c), shows that the posterior outcome distribution is robust to changes in the CACT prior.

Fig. 43 shows the robustness plane for the post-trial simulation, as derived in Sec. 5.6.2. The results could be used in two ways:

First, when planning a follow-up to the current trial, recruitment efforts and resource allocation can be focused on regions of low robustness. Second, the robustness plane could be utilized for personalization of treatments. For example, for new patients corresponding to regions of high robustness, the outcome of the CACT could be used to determine whether a patient should receive the device from one manufacturer or another.

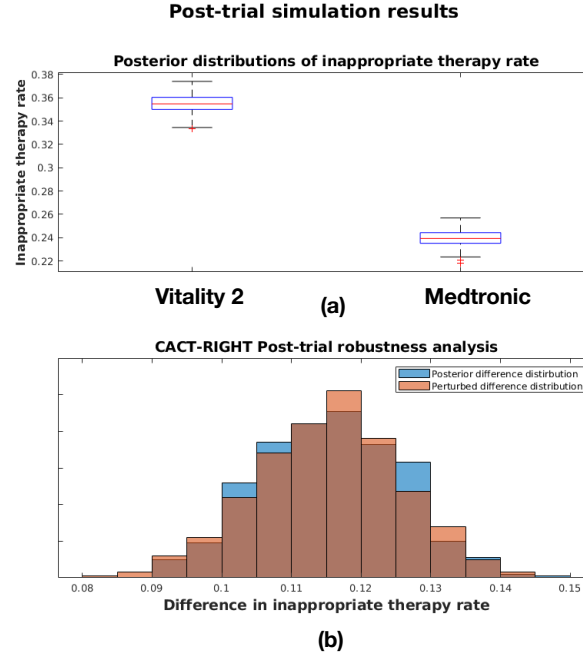


Figure 42: Results of post-trial simulation and robustness evaluation for CACT-RIGHT. (a) Estimated inappropriate therapy rate for each device. The difference is in line with actual RIGHT (b) Post-trial robustness analysis. Similar to Fig. 39, the CACT outcome demonstrates high robustness.

5.6.5. CACT vs other standard approaches

In order to assess the benefits of a CACT, three methods were compared: A frequentist approach comparing only proportions, a Bayesian historical controlled trial approach using the power prior method, and the CACT. Cohorts of endpoints were generated repeatedly according to the inappropriate therapy rate reported for RIGHT.

While varying the sample size, the power was estimated by determining the proportion of times each method correctly detected the relative difference in performance between the two devices. Fig. 44 shows how as the sample size increases, the power of the CACT increases much more rapidly compared to the other approaches. CACTs consistently had more power and achieved an average 55% improvement in power when compared to historical controlled trials. This implies that a CACT could achieve the same power as standard approaches using a smaller cohort, which could lead to a significant saving in costs due to the reduction in cohort size. For RIGHT, with the CACT, a comparable power could be achieved with

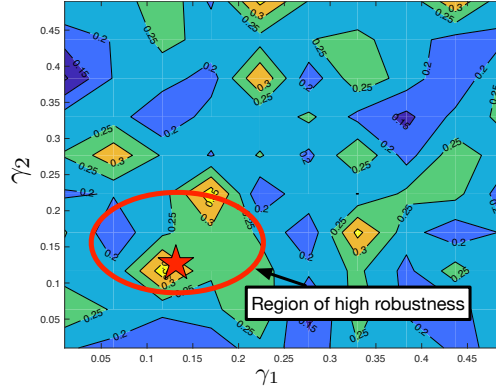


Figure 43: Post-trial analysis robustness plane. Subsequent trial may focus on areas of low robustness.

reduction in sample size up to 67% in comparison to historical controlled trials.

Fig. 45 shows the results of varying the underlying 'true' difference in inappropriate therapy rate with fixed cohort size of 400. As before, the power curve of the CACT increases rapidly compared to the other methods, further demonstrating the advantages of the CACT.

5.7. Discussion

5.7.1. Parameter selection for simulation

In CACT-RIGHT, the parameters for the Weibull discount need to be selected (see Sec. 5.6.3, (5.32)). The detailed procedure for evaluating the Type I error and power can be found in (Haddad, Himes, Thompson, et al., 2017). For CACT-RIGHT, a range of 'true' values for the parameter of interest, the difference in inappropriate therapy, is assumed based on information from clinical trials. For each of the true values, a cohort of endpoints was generated according to the that value. The values for κ and λ in (5.32) were set to a value within the range of $0.02 < \kappa < 1$ and $0.5 < \lambda < 1$, similar to (Haddad, Himes, Thompson, et al., 2017). For each value of κ and λ , numerous 'mock' trials were run and the type I errors and type II errors were tabulated. Note, here, we consider a type I error as when the true value was within the range of the null hypothesis, but the CACT concluded the opposite. Type II errors are defined similarly.

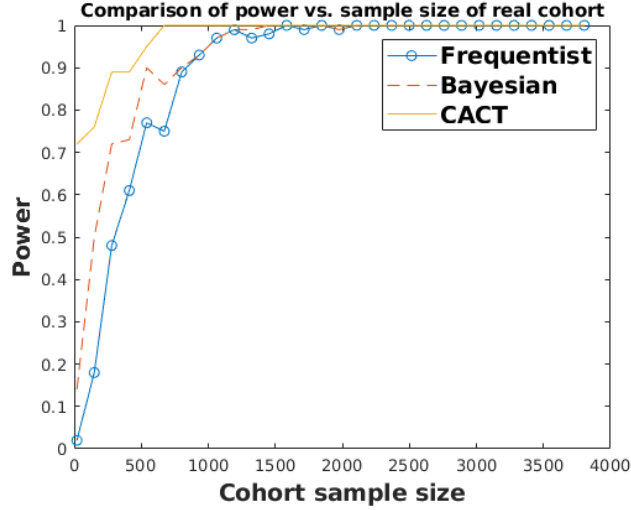


Figure 44: Comparison of CACT, Bayesian historical controlled trial and frequentist approach in terms of power vs. sample size of RIGHT. CACT requires a smaller cohort compared to the other methods for the same power. The CACT achieved an average 55.% improvement in power compared to historical controlled trials and reduced the sample size by approx. 67.%.

Fig. 46 shows an example of the Type I error and power over the range of κ and λ when a difference of 10% in the inappropriate therapy rate in favor of Vitality II is assumed. For CACT-RIGHT $\kappa = 0.05$ and $\lambda = 3$ was chosen for a power and type I error of 80% and 0.05%, respectively.

A similar procedure is used to determine the maximum effective sample size of the virtual cohort, $n_{0,max}$.

5.7.2. Steps towards general medical device evaluation

Up until now, the focus of this thesis was on the generation and incorporation of virtual cohorts and the systematic analysis of uncertainty from a Bayesian perspective. Therefore, the evaluation of the CACT was limited to the most basic form a clinical trial. However, it should be possible to use virtual cohorts in alternative clinical trial designs. An example of this would be in adaptive designs and factorial designs. In such cases, the evaluation with a virtual cohort can proceed in parallel with the various stages of the trial and act as a method to evaluate and determine the progression of trial stages.

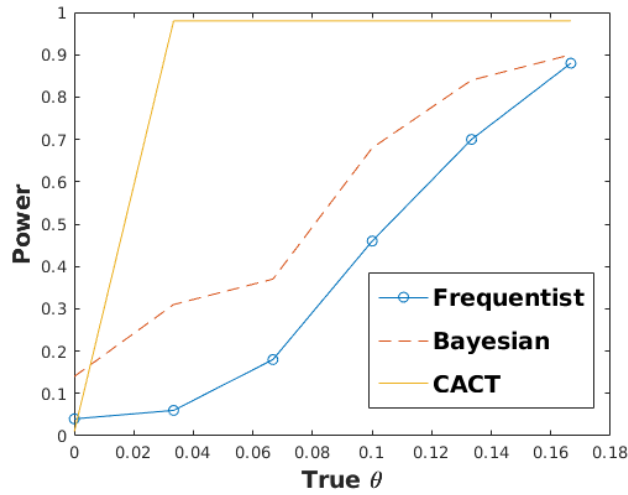


Figure 45: Comparison of CACT, Bayesian historical controlled trial and frequentist approach in terms of power vs true θ . In the positive θ direction, the CACT increases in power more rapidly than other methods.

The virtual cohort generation process and the framework of the CACT, though general in its definition, was shown concretely for the case of the ICD and the evaluation of the ICD discrimination algorithm. Application of the process to other medical devices, such as automated insulin pumps or fluid resuscitation systems would further support the case for using a CACT. When applying a CACT to these systems, determining the proper evaluation procedure will be important as the physiological simulators operate in a closed-loop manner. Different sources of uncertainty would also need to be identified.

Eventually an actual clinical trial would need to be executed in parallel with a CACT in order to validate the results of using a CACT. This would require additional collaboration across industry, academia, and regulatory institutions. Providing sufficient guidelines for regulation of such a trial is needed.

Similarly, there is potential for developing newer clinical trial designs with the goal of strengthening the credibility of a CACT. For example, the current results were derived through a retrospective analysis of RIGHT whose endpoint was the time-to-first inappropriate therapy. A future trial incorporating CACTs could be focused on endpoints which

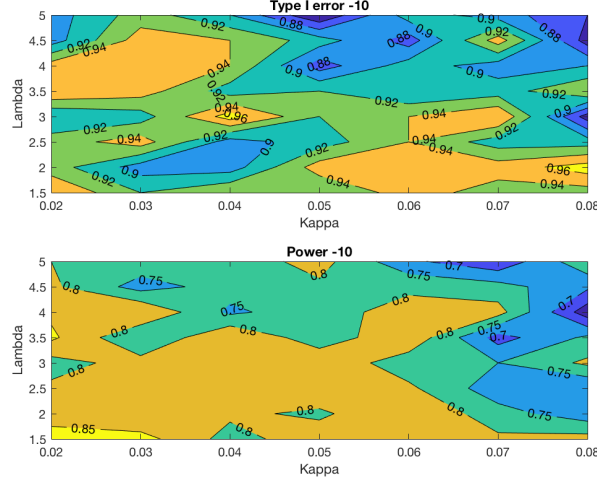


Figure 46: Results of parameter search for the Weibull discount function.

can be generated by a simulator. This would take further advantage of the power of CACTs while increasing the credibility of the outcomes.

Finally, from a more fundamental standpoint, developing mathematical properties which would guarantee the validity of the outcome of a simulator can also be explored. Such a result would allow for simulated data to be more readily combined with real data. Metrics like the proposed δ -robustness can be considered a step towards exploring such possibilities.

5.8. Chapter Conclusion

In this chapter, we first addressed the challenge of combining virtual cohorts with real data in the context of a clinical trial. Additionally, we addressed the challenge of considering uncertainty in the outcomes of a CACT due to the use of virtual cohorts. We provided a Bayesian statistical framework for modeling a CACT and its outcome. The framework incorporates results from simulated endpoints through the *pre-clinical prior* and explicitly models simulation assumptions. We then defined the δ -robustness of a CACT to quantify the uncertainty in the CACT outcome due to simulation assumptions and present a method for estimating the value. By applying the framework to the Rhythm ID Goes Head-to-head Trial (RIGHT), which evaluated the relative safety and efficacy of two software algorithms

in implantable cardiac devices, we confirmed that the results of the CACT aligned with the original trial. Furthermore, we demonstrated that using the framework increases the power of a clinical trial for a given sample size when compared to historical controlled trials and standard clinical trials. Finally, we demonstrated how the δ -robustness value can be used in analyzing the outcomes of a CACT through the δ -robustness plane.

CHAPTER 6 : A Data-driven Approach for Improving ICD Performance

6.1. Overview

In this chapter, we shift perspectives and address the final challenge of improving the performance of the device algorithm using a data-driven approach. This work was done in collaboration with Renukanandan Tumu, Nicole Chiou, Eric Eaton and Rahul Mangharam.

6.2. Introduction

As noted in previous chapters, ICDs have been shown to provide a significant reduction in mortality for patients undergoing life-threatening ventricular tachycardia. However, a major drawback of the device is the delivery of shocks for causes other than VT. These inappropriate therapies decrease the quality of life for patients undergoing ICD treatment as they remain in constant fear of triggering an inappropriate response from the device. Not only is the psychological burden increased, but these inappropriate therapy shocks are associated with increased mortality where even one inappropriate shock can increase the risk of mortality by 140% (Rees et al., 2011).

6.2.1. Manual configuration of current device algorithms and subpopulation exclusion

Current of ICD discrimination algorithms are structured with programmable parameter settings. As shown in Fig. 47, ICD algorithms, such as the Rhythm ID algorithm, are composed of rule-based discriminators with programmable rate thresholds and durations. The fixed discriminator components of the Rhythm ID algorithm require the setting of 30 tunable parameters - 21 for sensing, 9 for detection. These device settings are determined by clinicians and adjusted through *manual* intervention during routine check-ups. Studies have shown that adjusting these detection parameters can lead to decreased rates of inappropriate therapy (Moss et al., 2012). However, this manual process of adjusting parameters to fit a specific patient can easily lead to sub-optimal performance of the device.

Furthermore, guidelines and nominal parameter settings established through clinical trials

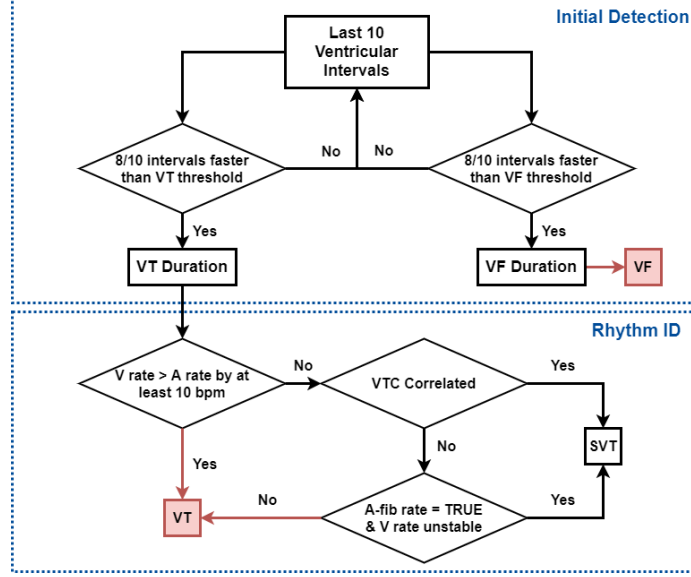


Figure 47: The Vitality 2, Rhythm ID detection pipeline with all discriminators implemented in this study. The squares represent classifications made by the detection algorithm, where the red boxes are rhythm classifications that trigger the application of therapy. (Figure credit: Nicole Chiou)

may result in poor performance for excluded, unobserved populations (Fig. 48). For example, it was found in (Garnreiter, 2017) that pediatric and congenital heart disease patient populations experienced inappropriate therapy complications at double the rate of the typical adult patient. This leads us to seek a better approach to improve ICD performance which would address both the manual tuning problem and subpopulation exclusion.

6.2.2. Motivation

The current process of device design and evaluation requires decades of additional research and clinical trials for each subpopulation and variation of the algorithm. As an alternative, we explore the potential for improving ICD performance by automatically obtaining a set of device parameters using data-driven approaches. This approach determines the optimal parameter setting of the algorithm based on observations in a given population. In particular, we demonstrate how Bayesian optimization can be used to improve the performance of the Rhythm ID algorithm. The results not only demonstrate the potential for automated improvement of ICD performance, but also highlights issues regarding optimization objectives and the concept of algorithmic fairness.

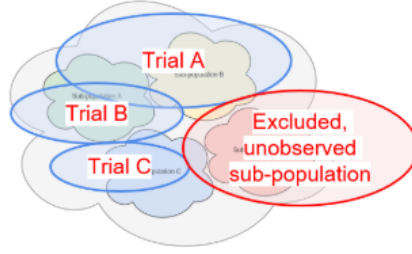


Figure 48: Specific subpopulations may be excluded from clinical trials, resulting in sub-optimal guidelines for those sub-populations.

6.3. Problem Formulation

We can formulate the problem of obtaining optimal parameters as a classification problem where we wish to minimize the empirical risk. For a data set of pair (X, Y) , where X is the physiological signal and Y is the label, the input to the device algorithm is the physiological signal and Y is the output. The input feature space $\mathcal{X} = \mathbb{R}^2$ has two channels where each one corresponds to a lead in the atrium and ventricle, respectively. The output label space is defined as follows,

$$Y = \begin{cases} 1 & \text{if the sample is a VT episode (VT or VF)} \\ 0 & \text{if the sample is not a VT episode (NSR, SVT)} \end{cases} \quad (6.1)$$

The device, e.g. ICD Discrimination algorithm, $f(X, \eta)$, outputs an estimate \hat{Y} .

We wish to find the optimal parameters

$$\eta^* = \arg \min_{\eta} \mathbb{E}_{(X, Y)} [\mathcal{L}(f(X, \eta), Y)] \quad (6.2)$$

where $\mathcal{L}(\cdot)$ is the empirical risk function.

In other words, we wish to find the parameters of the discrimination algorithm which minimizes the error rate. For the purposes of this chapter, our objective will be to optimize the accuracy. However, this could easily be adjusted to be some other objective, such as an

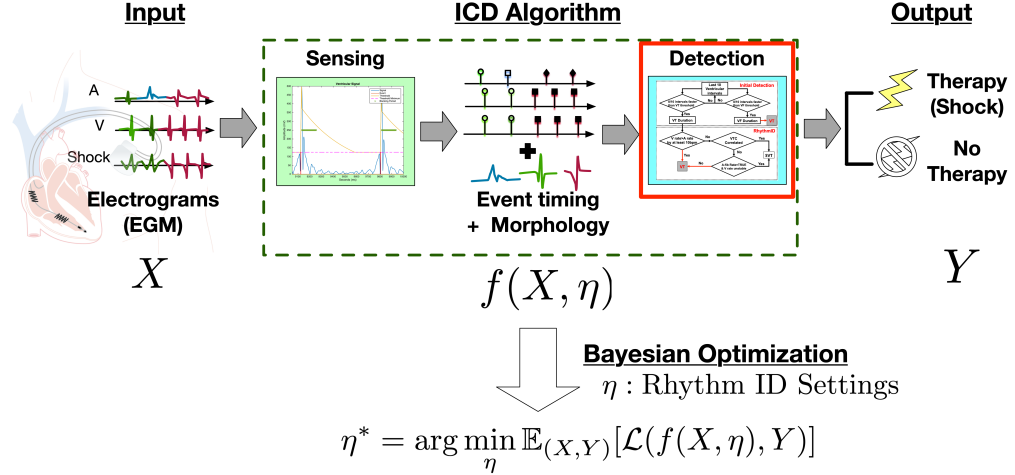


Figure 49: Data-driven device improvement with Bayesian Optimization. We seek to obtain parameter settings of the device algorithm which will improve performance through an automated, scalable approach that utilizes observation in the target population.

F1 objective.

6.4. Approach

In order to obtain these optimal parameters, we utilize Bayesian optimization. Bayesian Optimization, or BayesOpt, is a gradient-free optimization method typically used to optimize objective functions that are difficult or expensive to evaluate. Of importance is the *gradient-free* aspect of the method as it allows for the optimization of functions which may not have an explicit gradient.

Bayesian Optimization works by first building a surrogate model for the objective it is trying to optimize. The algorithm then quantifies the uncertainty in that surrogate using Gaussian process regression (GPR), a non-parametric approach that can provide uncertainty measurements. It uses this uncertainty estimate to construct an acquisition function to decide where in the parameter space to sample next, with the overall goal of estimating the best parameter settings to improve the value of the objective function.

A detailed background and additional resources regarding Bayesian optimization can be found in (Frazier, 2018). For this study, we used the implementation provided in BoTorch

Table 4: Default detection parameters for the Boston Scientific ICD Pipeline. The search space for Bayesian optimization in this study consist of the parameters in bold. (Ellenbogen et al., 2017; Gold, Ahmad, et al., 2012; Stroobandt, Barold, and Sinnaeve, 2009; Zanker et al., 2016). (Credit: Nicole Chiou)

Parameter Name	Function	Parameter Value
Post-ventricular atrial blanking (PVAB) period	Temporarily disables sensing in the A-channel after a ventricular sensed event	50 ms
Post-atrial ventricular blanking (PAVB) period	Temporarily disables sensing in the V-channel after an atrial sensed event	40 ms
Post-ventricular atrial refractory period (PVARP)	Records sensed events in the A-channel following a ventricular sensed event, but no actions are taken by the device	200 ms
Ventricular fibrillation (VF) threshold	Intervals below this threshold are classified as "fast" and contribute to the VF zone's detection window	320 ms
Ventricular tachycardia (VT) threshold	Intervals below this threshold are classified as "fast" and contribute to the VT zone's detection window	400 ms
Atrial fibrillation (AF) threshold	Indicates AF to be present if 6 out of 10 of the most recent A-intervals are below this threshold	200 ms
VF duration length	Duration that a rhythm must be sustained in the VF zone before applying therapy	10 sec
VT duration length	Duration that a rhythm must be sustained in the VT zone before applying therapy	10 sec

(Balandat et al., 2020). Bayesian optimization is best used for the optimization of twenty or fewer parameters and is often used to tune hyperparameters in machine learning.

Therefore in this study, we reduced the search space and optimize three detection parameters related to the interval length between beats: the VT threshold, VF threshold, and AF threshold. Some standard parameters and the default values of the discrimination algorithm are shown in Table 4. The VT and VF thresholds are part of the standard two-zone programming for ICDs and dictate the start of a VT and VF zone duration respectively. The decision to apply therapy is made at the end of the zone duration. If the tachycardia terminates before the end of the zone duration, then therapy is not applied. The AF threshold is used to determine whether an episode contains AF. The Rhythm ID detection pipeline relies heavily on rate-based information as the main portion of the algorithm and a large proportion of conditions can be decided solely based on this information. Future

Table 5: Class distribution for the pre-processed AAEL data for both the full-length data and a variant that 2343 down-sampled and segmented into five-second intervals for input into the neural networks.

Rhythm class	Number of episodes	Number of training examples
NSR	984	4840
SVT	964	1594
VT	566	2904

studies could include a larger search space which optimizes other parameters. For the optimization, the VF threshold was constrained to be less than the VT threshold since VF has a faster rate than VT and thus has a shorter peak-to-peak interval.

6.5. Dataset Description

The Ann Arbor Electrogram Libraries (AAEL) database used in Chapter 4 to create a dataset of sample rhythms and their corresponding label. This database contains EGM recordings collected from electrophysiological studies and contains a variety of cardiac arrhythmias as well as normal sinus rhythms (J. M. Jenkins and R. E. Jenkins, 2003) for 104 patients. From the database, the recording of same 62 patients were used to extract a training set and 42 patients were used to extract a validation set. This data was the same data used to extract the morphology of EGM waveforms in prior chapters.

The data are sampled at 1 kilohertz (kHz) and include three channels of recorded signals: bipolar ventricular, bipolar atrial, and unipolar ventricular.

For the current study, the portions of the signal which consisted of artifacts due to the acquisition process were manually adjudicated and removed. Table 5 shows the distribution of the episodes which were extracted from the dataset.

6.5.1. Data pre-processing

Each recorded signal contains segments of different rhythms at various lengths. These signals were annotated and separated into episodes by a cardiologist to isolate parts of the signals pertaining to the cardiac arrhythmias of interest (NSR/SVT/VT). In this study,

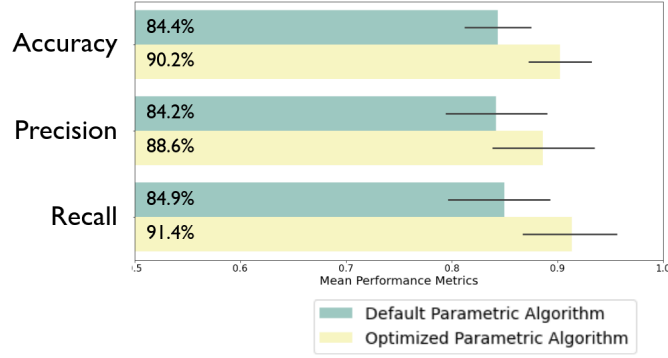


Figure 50: Barplot comparing the average summary metrics for the ICD algorithm with default settings (green) to that with parameter settings obtained with Bayesian optimization (yellow). The error bars represent 95% confidence intervals for the mean metrics.

VT and VF were categorized under the rhythm class VT since they both require therapy. The episodes were labeled according to the presence or absence of VT. For example, in the case that an episode contained both SVT and VT, the episode was labeled as VT. This is reasonable considering that the ICD prioritizes the treatment of VT with therapy.

6.6. Evaluation

For the evaluation, the data was divided into a training set and a validation set. Two different benchmarks were compared:

1. ICD Discrimination performance with default parameter settings
2. ICD Discrimination performance with parameters optimized with BayesOpt

For performance, the accuracy, recall, and specificity were computed based off the classification results on the validation set. For the default parameters, additional training was not needed and so only the results on the validation set were computed. A noisy Expected Improvement acquisition function was used and the optimization objective was set to improve the F1 score of classification.

Table 6: Default and optimized parameter settings for the detection rate thresholds.

	VF threshold	VT threshold	AF threshold
Default parameters	320	400	200
Optimized parameters	344	440	162

6.7. Results and Discussion

6.7.1. Results on the general population

Overall, the parameters obtained through Bayesian optimization with the accuracy objective improved the performance of the ICD algorithm when applied to the general population. Increases in the accuracy and recall were observed while maintaining the precision. The accuracy, precision, and recall metrics were computed for each of the patients in the validation set and averaged across all patients. The results are displayed in Figure 50. The accuracy, precision, and recall improved by 5.8%, 4.4% , and 6.5%, respectively.

Overall, the optimized parameters resulted in an improved performance for all patients. The increase in accuracy is mainly due to the change in the rate threshold parameter settings. As shown in Table 6, the VF threshold was set to be more strict (shorter interval) after optimization while the VT threshold became more flexible (longer interval). These changes allow for more episodes of VT to trigger the start of a VT duration and maximize the usage of other discriminators in the decision-making pipeline.

These results were significant in that they aligned with the MADIT-RIT clinical trial (Moss et al., 2012). In this clinical trial, increasing the thresholds and durations of discrimination algorithms resulted in reduced inappropriate therapy and mortality. Moreover, as noted in Chapter 4, during the in-silico evaluation of ICD device algorithms shown in Fig. 27, it was demonstrated that higher thresholds resulted in a decreased rate of inappropriate therapy. These results not only demonstrate the feasibility of utilizing a data-driven approach for obtaining optimal parameters, it also further strengthens the case for utilizing computer-aided clinical trials in device performance evaluation.

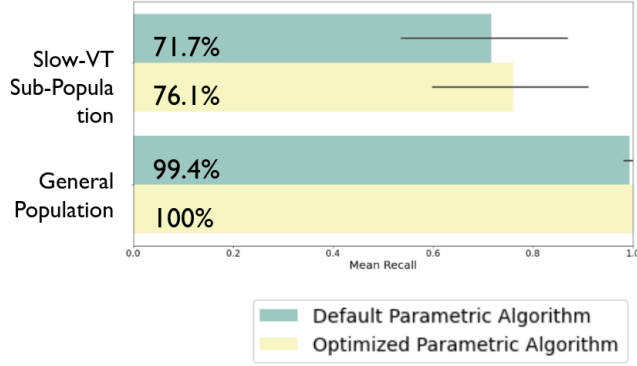


Figure 51: Barplot comparing the average recall for the slow-VT sub-population and the general population under the default parametric settings (green) and the optimized parametric settings (yellow), with error bars depicting the 95% confidence intervals.

6.7.2. Limitations on slow-VT population

An interesting observation is that despite the overall performance gain for the general population, for the slow-VT subpopulation, the performance benefit was not observed. Here, slow-VT is defined as VT with ventricular rate greater than 400ms. As shown in Fig. 51, in both cases with the default algorithm settings and with parameters optimized with Bayesian Optimization, the performance degraded, yielding poor recall for these slow-VT patients when compared to the remainder of the general population. This result highlights the need for understanding how the definition of the objective function can influence the outcome of optimization. In principle, the optimization algorithm did what it was designed to do, but due to the definition of the objective, certain subpopulations did not benefit. Understanding the trade-offs and defining an objective that will improve all subpopulations relates to algorithmic fairness (Cerbone, 2021) and should be pursued in future research.

6.8. Chapter Conclusion

In this chapter we approached the problem of improving the discrimination algorithm itself. For this, we utilized a data-driven approach where optimized parameters were found according to observations in a dataset using Bayesian optimization. We were able to find parameters for the algorithm which improved the accuracy, precision, and recall by 5.8%, 4.4% , and 6.5%, respectively, when compared to the default settings of the device.

However, trade-offs exist in both the selection of the objective to be optimized, as well as issues regarding the benefits to particular sub-population. It is possible that alternative data-driven approaches can improve such issues, but even in those cases defining the proper objective that can accurately reflect the requirements of the device to provide benefit across all segments of the target population will be needed. Virtual cohort generation and CACTs can play an important role in identifying subpopulations with degraded performance and evaluating algorithms that address issues of fairness.

CHAPTER 7 : Conclusion

7.1. Summary of Contributions

In this thesis, we have explored the problem of improving the evaluation of medical devices in the context of the clinical trial. Specifically, we sought to improve the power of a trial by utilizing a virtual cohort of endpoints generated using physiological simulation. In addition, we aimed to improve device performance with an automated approach to device configuration.

Generating a virtual cohort for use in the context of a clinical trial

In order to generate a virtual cohort in the context of a clinical trial, we discussed the necessity and requirements of physiological models and demonstrated the principles through by proposing an EGM simulator for evaluation of ICDs. Within the context of evaluating the ICD, the process of virtual cohort generation was presented and improved through the use of Bayesian hierarchical models. The effectiveness was demonstrated empirically through the case study of RIGHT, an actual clinical trial for ICD performance evaluation. In this case study, the usage of Bayesian hierarchical models and incorporating historical information resulted in an average increase in accuracy of predicted inappropriate therapy rates by approximately 42%.

Computer-aided clinical trials to improve statistical power with virtual cohorts

Next, as a step towards regulatory acceptance of simulation results, the uncertainty resulting from assumptions needed to be considered during evaluation. For this, we proposed a Bayesian statistical framework for incorporating virtual cohorts which we call the Computer-aided Clinical Trial (CACT). The framework explicitly models sources of uncertainty in simulated outcomes and incorporates a mechanism for combining both simulated and real data. This allowed for historical information to be incorporated in the evaluation process in a systematic manner and enabled the combination of virtual cohorts with real data which would be collected during a clinical trial. Thus, the framework encompasses all stages of a

clinical trial, from the pre-clinical design stage, interim analysis, and trial conclusion.

Through the case study of the RIGHT trial and ICD evaluation, we demonstrated the clear benefit of the CACT and virtual cohorts in terms of the power vs. sample size. We showed empirically that for RIGHT, a comparable power could be achieved with a sample size reduced by up to 67%. Moreover, we were able to make these conclusions through a simulation study which lasted roughly 5 weeks compared to the actual trial which lasted 5 years. The savings in terms of time, money, and human effort are incomparable.

Uncertainty quantification of computer-aided clinical trial outcomes

To analyze the uncertainty in the simulation outcomes, we introduced the δ -robustness metric and demonstrated the utility of the metric through the robustness evaluation of CACT outcomes. The δ -robustness metric and robustness evaluation can be used to determine specific subpopulations where there is a large degree of uncertainty in the CACT outcome. These results can possibly be utilized in the design of a clinical trial in order to more efficiently allocate limited resources to further improve the power of a trial.

Device performance improvement with data-driven methods

Finally, our approach based on Bayesian optimization was used to improve the performance of the ICD algorithm found in Boston Scientific ICDs. Through an automated, data-driven procedure, we obtained parameter settings for the algorithm which improved the accuracy, precision, and recall by 5.8%, 4.4% , and 6.5%, respectively, compared to the device performance with default settings.

7.2. Open Questions

Investigating the problem of medical device evaluation with virtual cohorts resulted in the contributions described above, but also led to the discovery of additional directions for research which we leave as open questions:

Necessity of regulatory support for CACT adoption

The results of this thesis clearly demonstrate the benefits of utilizing simulated data for the evaluation of medical devices. Though the results were presented within the application to the ICD, repeating the process for other medical devices will most likely result consists of following the procedure outlined in this thesis. However, a fundamental issue is the need for widespread collaboration between academia, industry, and regulatory officials to accept and promote the usage of simulation-based regulatory evaluation. Initiatives already exist, such as the ASME V&V 40 (ASME, 2018), to define and regulate what can be accepted in terms of simulation-based evidence for medical devices. This is an ongoing effort and resolving this issue will bring exciting developments in the future of medical device evaluation.

CACTs for the design and evaluation of medical devices to improve fairness

In the final chapter of this thesis, we explored the problem of data-driven methods for improving the performance of medical devices. There are clear benefits to using more data-driven methods, such as machine learning, in order to adjust to changing conditions and address the heterogeneity in the population. As device complexity increases and more data-driven components are incorporated into device algorithms, conventional evaluation through clinical trials will be less effective and simulation-based evaluation, such as with the CACT, will occupy a larger portion of the evaluation process.

Whether it is to improve the efficiency of evaluation or to improve the performance of the device, an important question to be addressed is how to resolve the problem of fairness in such data-driven models. As was observed in the final study, even though the overall performance on a target population may increase, specific subpopulations may not see those benefits and in the worst case, show degraded performance. The issue of fairness is not limited to the medical domain and is becoming a prominent issue within the scientific community (Cerbone, 2021; Dwork et al., 2011). Exploring the issues of fairness within the medical domain, especially in the context of defining improvement for a medical device which uses data-driven models will be an interesting direction for future research.

Beyond applications in medical devices

The problem of evaluating and improving systems using alternative, simulation-based data is a broad topic relevant beyond medical devices. Off-policy learning in reinforcement learning and bridging the sim-to-real gap are topics where both improving the model and evaluation of the performance require sources of data which are separate from the target domain. A related topic is in transfer learning, where the goal is to maintain or improve performance on a target domain task by a model which was trained on a source domain. The principles of clinical trials and hypothesis testing can be applied to such problems in order to evaluate generalization in such problem settings. Moreover, CACTs and uncertainty quantification with δ -robustness could potentially be expanded to such applications and provide insight into the development of such systems. For example, in autonomous driving, simulation environments of vehicles and traffic have been used to accelerate safety evaluation of autonomous vehicles (O’Kelly et al., 2018). In this scenario, combining simulated data with real data collected during driving in order to both evaluate and improve the performance of the vehicle would vastly improve the efficiency of development. Scenarios which require additional data could be identified using the process of CACTs and robustness evaluation. In this way, expanding the frontiers of CACTs beyond the domain of medical devices could potentially lead to interesting discoveries in the field of machine learning and artificial intelligence.

7.3. Ending Remarks

Clinical trials and hypothesis testing are valuable tools in the development and evaluation of medical treatments and therapies. The contributions of this thesis provide a step towards improving such methods in the domain of medicine and opens the potential to be applied more generally in domains beyond medicine.

APPENDIX

A.1. Hybrid systems and simulations

Definition 1. A *hybrid automaton* is a tuple $\mathcal{H} = (X, L, H_0, \{f_\ell\}, \text{Inv}, E, \{R_{ij}\}_{(i,j) \in E}, \{G_{ij}\}_{(i,j) \in E})$ where $X \subset \mathbb{R}^n$ is the continuous state space equipped with the Euclidian norm $\|\cdot\|$, $L \subset \mathbb{N}$ is a finite set of modes, $H_0 \subset X \times L$ is an initial set, $\{f_\ell\}_{\ell \in L}$ determine the continuous evolutions with unique solutions, $\text{Inv} : L \rightarrow 2^X$ defines the invariants for every mode, $E \subset L^2$ is a set of discrete transitions, $G_{ij} \subset X$ is guard set for the transitions (so \mathcal{H} transitions $i \rightarrow j$ when $x \in G_{ij}$), $R_{ij} : X \rightarrow X$ is an edge-specific reset function.

Set $H = L \times X$. Given $(\ell, x_0) \in H$, the *flow* $\theta_\ell(\cdot; x_0) : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ is the solution to the IVP $\dot{x}(t) = f_\ell(x(t))$, $x(0) = x_0$.

The associated *transition system* is $T_{\mathcal{H}} = (H, E \cup \{\tau\}, \rightarrow, H_0)$ where H is the state set, $E \cup \{\tau\}$ is the label set for transitions, H_0 is the set of initial states, and $\rightarrow = (\bigcup_{e \in E} \xrightarrow{e}) \cup \xrightarrow{\tau}$ where $(i, x) \xrightarrow{e} (j, y)$ iff $e = (i, j)$, $x \in G_{ij}$, $y = R_{ij}(x)$ and $(i, x) \xrightarrow{\tau} (j, y)$ iff $i = j$ and there exists a flow $\theta_i(\cdot; x)$ of \mathcal{H} and $t \geq 0$ s.t. $\theta_i(t; x) = y$ and $\forall t' \leq t$, $\theta_i(t'; x) \in \text{Inv}(i)$. Let \sim be an equivalence relation on H and H/\sim the corresponding partition. Let $\mathcal{F}_t(H/\sim)$ be the coarsest bisimulation with respect to $\xrightarrow{\tau}^1$ respecting the partition H/\sim , and $\mathcal{F}_d(H/\sim) := \{(h_1, h_2) \mid (h_1 \xrightarrow{e} h'_1) \implies (\exists e' \in E, h'_2 . h_2 \xrightarrow{e'} h'_2 \wedge h'_1 \sim h'_2)\} \cap H/\sim$ Vladimerou et al., 2008. The iteration

$$W_0 = \mathcal{F}_t(H/\sim), \quad \forall i \geq 0, \quad W_{i+1} = \mathcal{F}_t(\dot{F}(W_i)) \quad (\text{A.1})$$

computes a bisimulation of \mathcal{H} . However it does not necessarily terminate for hybrid systems because the system's reach set might intersect a given block of H/\sim an infinite number of times (see Lafferriere, Pappas, and Sastry, 2000 for an example). The class of systems introduced in the next section has the property that the iteration does terminate for it and returns a finite \mathcal{S} .

¹I.e., \mathcal{F}_t only considers the continuous transition relation: it is a bisimulation of $T_{\mathcal{H}}^c := (H/\sim, \{*\}, \xrightarrow{\tau}, H_0/\sim)$.

Given a set of atomic propositions, if \sim is s.t. $\eta \sim \eta'$ iff both states satisfy exactly the same atomic propositions, then model checking temporal logic properties can be done on the finite bisimulation instead of the possibly infinite \mathcal{H} .

A.2. O-minimality and STORMED systems

We give a very brief introduction to o-minimal structures. A more detailed introduction can be found in Lafferriere, Pappas, and Sastry, 2000 and references therein. We are interested in sets and functions in \mathbb{R}^n that enjoy certain finiteness properties, called order-minimal sets (o-minimal). These are defined inside *structures* $\mathcal{A} = (\mathbb{R}, <, +, -, \cdot, \exp, \dots)$. The subsets $Y \subset \mathbb{R}^n$ we are interested in are those that are *definable* using first-order formulas φ : $Y = \{(a_1, \dots, a_n) \in \mathbb{R}^n \mid \varphi(a_1, \dots, a_n)\}$. (First-order formulas use the boolean connectives and the quantifiers \exists, \forall). The atomic propositions from which the formulas are recursively built allow only the operations of the structure \mathcal{A} on the real variables and constants, and the relations of \mathcal{A} and equality. For example $2x - 3.6y < 3z$ and $x = y$ are valid atomic propositions of the structure $\mathcal{L}_{\mathbb{R}} = (\mathbb{R}, <, +, -, \cdot)$, while $\cosh(x) < 3z$ is not because \cosh is not in the structure. These structures are already sufficient to describe a set of dynamics rich enough for our purposes and for various classes of linear systems.

Definition 2. A theory of (\mathbb{R}, \dots) is *o-minimal* if the only definable subsets of \mathbb{R} are finite unions of points and (possibly unbounded) intervals. A function $f : x \mapsto f(x)$ is o-minimal if its graph $\{(x, y) \mid y = f(x)\}$ is a definable set.

We use the terms o-minimal and definable interchangeably, and they refer to $\mathcal{L}_{\text{exp}} = (\mathbb{R}, <, +, -, \cdot, \exp)$ which is known to be o-minimal. The dot product between $x, y \in \mathbb{R}^n$ is denoted $x \cdot y$, and $d(Y, S) = \inf\{\|y - s\| \mid (y, s) \in Y \times S\}$.

Definition 3. Vladimerou et al., 2008. A *STORMED hybrid system* (SHS) Σ is a tuple $(\mathcal{H}, \mathcal{A}, \phi, b_-, b_+, d_{\min}, \epsilon, \zeta)$ where \mathcal{H} is a hybrid automaton, \mathcal{A} is an o-minimal structure, $d_{\min}, \epsilon, \zeta$ are positive reals, $b_-, b_+ \in \mathbb{R}$ and $\phi \in X$ such that:

(S) The system is d_{min} -separable,

meaning that for any $e = (\ell, \ell') \in E$ and $\ell'' \neq \ell', d(R_e(G_{(\ell, \ell')}), G_{(\ell', \ell'')}) > d_{min}$ ²

(T) The flows (i.e., the solutions of the ODEs) are Time-Independent with the Semi-Group property (TISG), meaning that for any $\ell \in L, x \in X$, the flow θ_ℓ starting at (ℓ, x) satisfies:

1) $\theta_\ell(0; x) = x$, 2) for every $t, t' \geq 0$, $\theta_\ell(t + t'; x) = \theta_\ell(t'; \theta_\ell(t; x))$

(O) All the sets and functions of \mathcal{H} are definable in the o-minimal structure \mathcal{A}

(RM) The resets and flows are monotonic with respect to the same vector ϕ , meaning that

1) (Flow monotonicity) for all $\ell \in L, x \in X$ and $t, \tau \geq 0$, $\phi \cdot (\theta_\ell(t + \tau; x) - \theta_\ell(t; x)) \geq \epsilon \|\theta_\ell(t + \tau; x) - \theta_\ell(t; x)\|$, and

2) (Reset monotonicity) for any edge $(\ell, \ell') \in E$ and any $x^-, x^+ \in X$ s.t. $x^+ = R_{\ell, \ell'}(x^-)$,

1. if $\ell = \ell'$, then either $x^- = x^+$ or $\phi \cdot (x^+ - x^-) \geq \zeta$

2. if $\ell \neq \ell'$, then $\phi \cdot (x^+ - x^-) \geq \epsilon \|x^+ - x^-\|$

(ED) Ends are Delimited: for all $e \in E$ we have $\phi \cdot x \in (b_-, b_+)$ for all $x \in G_e$

Intuitively, the above conditions imply the trajectories of the system always move a minimum distance along ϕ whether flowing or jumping, which guarantees that no area of the state space will be visited infinitely often. This is at the root of the finiteness properties of STORMED systems.

A.3. Numerical computation of ϵ perturbations for discrete distributions

We solve the following feasibility problem to find the parameters $\alpha \in \mathbb{R}^n$ of a distribution that is ϵ Hellinger distance away from a given distribution with parameters $\beta \in \mathbb{R}^n$. Let $\gamma \in \mathbb{R}^n$ be such that $\gamma_i = \sqrt{\alpha_i} \forall i = 1, \dots, n$. The (non-convex) feasibility problem then becomes:

²The original definition of separability Vladimerou et al., 2008 required the guards themselves to be separated, which is insufficient to guarantee that if \mathcal{H} flows, it flows a uniform minimum distance along ϕ . Indeed assume the guards are separated. If $x \in G_{(\ell, \ell')}$ and $y = R_{\ell, \ell'}(x)$, it can be that $y \in G_{(\ell', \ell'')}$ and thus a jump happens, even though $G_{(\ell, \ell')}$ and $G_{(\ell', \ell'')}$ are separated. Therefore we need $d(y, G_{(\ell', \ell'')}) > d_{min}$ for all $y \in R_e(G_e)$, which is the condition we use in Def. 3. The properties of SHS, in particular the existence of finite bisimulation, are therefore preserved by this change.

minimize $_{\gamma}$ 0, subject to,

$$\mathbf{1} \succcurlyeq \gamma \succcurlyeq \mathbf{0} \tag{A.2a}$$

$$\gamma^T \gamma = 1 \tag{A.2b}$$

$$\gamma^T \gamma - \begin{bmatrix} 2\sqrt{\beta_1} & 2\sqrt{\beta_2} & \dots & 2\sqrt{\beta_n} \end{bmatrix} \gamma + \mathbf{1}^T \beta - 2\epsilon^2 = 0 \tag{A.2c}$$

Here, $\mathbf{1} \in \mathbb{R}^n$ (and $\mathbf{0}$) is a vector with all elements being 1 (and 0). Eqs. A.2a and A.2b ensure that $\alpha_i = \gamma_i^2 \forall i$ form the parameters of a probability distribution, i.e. are bounded between 0 and 1, and sum up to 1 (respectively). Eq. A.2c enforces the condition that the square of the Hellinger distance between the distributions given by α and β equals ϵ^2 . This follows directly from Def. 12.

Note that despite the constraints being linear and quadratic in the variable γ , the resulting optimization is non-convex because of the Eq. A.2c, which is a quadratic equality constraint. However the problem can still be quickly solved by off-the-shelf non-convex optimization algorithms due to the dimension of the variable n being small. Our implementation of this feasibility problem in MATLAB is one such instance of this.

BIBLIOGRAPHY

- Abbas, Houssam, Kuk Jin Jang, Zhihao Jiang, and Rahul Mangharam (Apr. 2016). “Towards Model Checking of Implantable Cardioverter Defibrillators”. In: *Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*. HSCC ’16. New York, NY, USA: Association for Computing Machinery, pp. 87–92. ISBN: 978-1-4503-3955-1. DOI: 10.1145/2883817.2883841.
- Abbas, Houssam, Kuk Jin Jang, and Rahul Mangharam (Feb. 2017). “Nonlinear Hybrid Automata Model of Excitable Cardiac Tissue”. en-US. In: *EPiC Series in Computing*. Vol. 43. ISSN: 2398-7340. EasyChair, pp. 1–8. DOI: 10.29007/5zfk.
- Abbas, Houssam, Zhihao Jiang, Kuk Jin Jang, Marco Beccani, Jackson Liang, and Rahul Mangharam (Oct. 2016). “High-level modeling for computer-aided clinical trials of medical devices”. In: *Proc. IEEE Int. High Level Design Validation and Test Workshop (HLDVT)*, pp. 85–92. DOI: 10.1109/HLDVT.2016.7748260.
- Abbas, Houssam, Hans Mittelmann, and Georgios Fainekos (2014). *Formal property verification in a conformance testing framework*.
- Abbas, Ismail (2016). “Modeling and Simulation in Clinical Trials Modeling and Simulation in Clinical Trials”. In: 6.April, pp. 2016–2018. DOI: 10.22360/SpringSim.2016.MSM.002.
- Alur, R., T.A. Henzinger, G. Lafferriere, and G.J. Pappas (July 2000). “Discrete abstractions of hybrid systems”. en. In: *Proceedings of the IEEE* 88.7, pp. 971–984. ISSN: 0018-9219, 1558-2256. DOI: 10.1109/5.871304.
- ASME (2018). *Assessing Credibility of Computational Modeling through Verification & Validation: Application to Medical Devices - ASME*. en.
- Balandat, Maximilian, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy (Dec. 2020). “BoTorch: A Framework for Effi-

- cient Monte-Carlo Bayesian Optimization”. In: *arXiv:1910.06403 [cs, math, stat]*. arXiv: 1910.06403.
- Berger, Ronald D et al. (2006). “The Rhythm ID Going Head to Head Trial (RIGHT): Design of a Randomized Trial Comparing Competitive Rhythm Discrimination Algorithms in Implantable Cardioverter Defibrillators”. In: *Journal of Cardiovascular Electrophysiology* 17.7, pp. 749–753. ISSN: 1540-8167. DOI: 10.1111/j.1540-8167.2006.00463.x.
- Bogomolov, Sergiy, Christian Schilling, Ezio Bartocci, Gregory Batt, Hui Kong, and Radu Grosu (2015). “11th International Haifa Verification Conference, HVC 2015, Proceedings”. In: ed. by Nir Piterman. Cham: Springer International Publishing. Chap. Abstraction-Based Parameter Synthesis for Multiaffine Systems, pp. 19–35. ISBN: 978-3-319-26287-1. DOI: 10.1007/978-3-319-26287-1_2.
- Boston Scientific Corporation (2007). “The Compass - Technical Guide to Boston Scientific Cardiac Rhythm Management Products”. In: *Device Documentation*.
- Box, George E. P. (Dec. 1976). “Science and Statistics”. In: *Journal of the American Statistical Association* 71.356. Publisher: Taylor & Francis, pp. 791–799. ISSN: 0162-1459. DOI: 10.1080/01621459.1976.10480949.
- Burns, Patricia B., Rod J. Rohrich, and Kevin C. Chung (July 2011). “The Levels of Evidence and their role in Evidence-Based Medicine”. In: *Plastic and reconstructive surgery* 128.1, pp. 305–310. ISSN: 0032-1052. DOI: 10.1097/PRS.0b013e318219c171.
- C. D. Swerdlow and others (2002). “Discrimination of Ventricular Tachycardia from Supraventricular Tachycardia by a Downloaded Wavelet Transform Morphology Algorithm: A Paradigm for Development of Implantable Cardioverter Defibrillator Detection Algorithms”. In: *J. Cardiovascular Electrophysiology* 13.

- C. Toffanin, M. Messori, F. Di Palma, G. De Nicolao, C. Cobelli, L. Magni (2014). “Artificial Pancreas: Model Predictive Control Design from Clinical Experience”. In: *J Diabetes Sci Technol* 7, pp. 1470–1483.
- Campbell, Gregory (2017). “Bayesian methods in clinical trials with applications to medical devices”. In: *Communications for Statistical Applications and Methods* 24.6, pp. 561–581.
- Cerbone, Henry (Oct. 2021). “Providing a Philosophical Critique and Guidance of Fairness Metrics”. In: *arXiv:2111.04417 [cs]*. arXiv: 2111.04417.
- Clayton, R. H., O. Bernus, E. M. Cherry, H. Dierckx, F. H. Fenton, L. Mirabella, A. V. Panfilov, F. B. Sachse, G. Seemann, and H. Zhang (Jan. 2011). “Models of cardiac tissue electrophysiology: Progress, challenges and open questions”. en. In: *Progress in Biophysics and Molecular Biology*. Cardiac Physiome project: Mathematical and Modelling Foundations 104.1, pp. 22–48. ISSN: 0079-6107. DOI: 10.1016/j.pbiomolbio.2010.05.008.
- Cobelli, C., E. Renard, and B. Kovatchev (2011). “Artificial pancreas: past, present, future”. In: *Diabetes*.
- Correa de Sa, Daniel D., Nathaniel Thompson, Justin Stinnett-Donnelly, Pierre Znojkwicz, Nicole Habel, Joachim G. Muller, Jason H.T. Bates, Jeffrey S. Buzas, and Peter S. Specter (Dec. 2011). “Electrogram Fractionation”. In: *Circ Arrhythm Electrophysiol* 55 (3), pp. 909–916. DOI: 10.1161/CIRCEP.111.965145.
- Daubert, James P. et al. (2008). “Inappropriate Implantable Cardioverter-Defibrillator Shocks in MADIT II: Frequency, Mechanisms, Predictors, and Survival Impact ”. In: *Journal of the American College of Cardiology* 51.14, pp. 1357–1365.
- Dubin, Jonathan R., Stephen D. Simon, Kirsten Norrell, Jacob Perera, Jacob Gowen, and Akin Cil (May 2021). “Risk of Recall Among Medical Devices Undergoing US Food and Drug Administration 510(k) Clearance and Premarket Approval, 2008-2017”. In: *JAMA*

- Network Open* 4.5, e217274. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.7274.
- Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel (Nov. 2011). “Fairness Through Awareness”. In: *arXiv:1104.3913 [cs]*. arXiv: 1104.3913.
- Ellenbogen, Kenneth A., Bruce L. Wilkoff, G. Neal Kay, Chu-Pak Lau, and Angelo Auricchio (2017). “Follow-Up and Programming”. In: *Clinical Cardiac Pacing, Defibrillation and Resynchronization Therapy*. Elsevier. ISBN: 978-0-323-37804-8.
- Emanuel, E. J., D. Wendler, and C. Grady (May 2000). “What makes clinical research ethical?” eng. In: *JAMA* 283.20, pp. 2701–2711. ISSN: 0098-7484. DOI: 10.1001/jama.283.20.2701.
- Ermentrout, G. Bard and Leah Edelstein-Keshet (1993). “Cellular Automata Approaches to Biological Modeling”. In: *Journal of Theoretical Biology* 160.1, pp. 97–133. ISSN: 0022-5193. DOI: <http://dx.doi.org/10.1006/jtbi.1993.1007>.
- Evans, Scott R. (Jan. 2010). “Fundamentals of clinical trial design”. In: *Journal of experimental stroke & translational medicine* 3.1, pp. 19–27. ISSN: 1939-067X.
- Faris, Owen and Jeffrey Shuren (Apr. 2017). “An FDA Viewpoint on Unique Considerations for Medical-Device Clinical Trials”. In: *New England Journal of Medicine* 376.14. Publisher: Massachusetts Medical Society _eprint: <https://doi.org/10.1056/NEJMr1512592>, pp. 1350–1357. ISSN: 0028-4793. DOI: 10.1056/NEJMr1512592.
- Farrar, John T., Andrea B. Troxel, Kevin Haynes, Ian Gilron, Robert D. Kerns, Nathaniel P. Katz, Bob A. Rappaport, Michael C. Rowbotham, Ann M. Tierney, Dennis C. Turk, and Robert H. Dworkin (Aug. 2014). “Effect of variability in the 7-day baseline pain diary on the assay sensitivity of neuropathic pain randomized clinical trials: an ACTION study”. eng. In: *Pain* 155.8, pp. 1622–1631. ISSN: 1872-6623. DOI: 10.1016/j.pain.2014.05.009.

- FDA, U.S. (Mar. 2019). *The 510(k) Program: Evaluating Substantial Equivalence in Pre-market Notifications [510(k)]*. en. Publisher: FDA.
- Fenton, F. and A. Karma (1998). “Vortex dynamics in three-dimensional continuous myocardium with fiber rotation: Filament instability and fibrillation”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 8.1, pp. 20–47.
- Fenton, F. H. and E. M. Cherry (2008). “Models of cardiac cell”. In: *Scholarpedia*. Vol. 3. 8, p. 1868.
- Fogel, David B. (Aug. 2018). “Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review”. In: *Contemporary Clinical Trials Communications* 11, pp. 156–164. ISSN: 2451-8654. DOI: 10.1016/j.conctc.2018.08.001.
- Fogoros (2017). “Electrophysiologic Testing”. In: John Wiley & Sons, Ltd. Chap. 1, pp. 1–12. ISBN: 9781119235767. DOI: <https://doi.org/10.1002/9781119235767.ch1>.
- Fontenla, Adolfo (2016). “Incidence of arrhythmias in a large cohort of patients with current implantable cardioverter-defibrillators in Spain: results from the UMBRELLA Registry”. In: *EP Europace* 18.11, pp. 1726–1734.
- Food, Drug Administration, et al. (2011). *Guidance for industry and FDA staff: Guidance for the use of bayesian statistics in medical device clinical trials*.
- Frazier, Peter I. (July 2018). “A Tutorial on Bayesian Optimization”. In: *arXiv:1807.02811 [cs, math, stat]*. arXiv: 1807.02811.
- Garnreiter, Jason M. (Nov. 2017). “Inappropriate ICD Shocks in Pediatric and Congenital Heart Disease Patients”. In: *The Journal of Innovations in Cardiac Rhythm Management* 8.11, pp. 2898–2906. ISSN: 2156-3977. DOI: 10.19102/icrm.2017.081104.

- Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin (2014). *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL.
- Giorgini, Paolo, Alan B. Weder, Elizabeth A. Jackson, and Robert D. Brook (Sept. 2014). “A review of blood pressure measurement protocols among hypertension trials: implications for ”evidence-based” clinical practice”. eng. In: *Journal of the American Society of Hypertension: JASH* 8.9, pp. 670–676. ISSN: 1878-7436. DOI: 10.1016/j.jash.2014.07.024.
- Giovagnoli, A and M Zagoraiou (2012). “Simulation of clinical trials: a review with emphasis on the design issues”. In: *Statistica* 1, pp. 63–80.
- Gold, Michael R., Saleem Ahmad, Kevin Browne, Kellie Chase Berg, Lisa Thackeray, and Ronald D. Berger (2012). “Prospective comparison of discrimination algorithms to prevent inappropriate ICD therapy: Primary results of the Rhythm ID Going Head to Head Trial ”. In: *Heart Rhythm* 9.3, pp. 370–377. ISSN: 1547-5271. DOI: <http://dx.doi.org/10.1016/j.hrthm.2011.10.004>.
- Gold, Michael R., Torsten Sommer, Juerg Schwitter, Ahmed Al Fagih, Timothy Albert, Béla Merkely, Michael Peterson, Allen Ciuffo, Sung Lee, Lynn Landborg, Jeffrey Cerkenvenik, Emanuel Kanal, and Evera MRI Study Investigators (June 2015). “Full-Body MRI in Patients With an Implantable Cardioverter-Defibrillator: Primary Results of a Randomized Study”. eng. In: *Journal of the American College of Cardiology* 65.24, pp. 2581–2588. ISSN: 1558-3597. DOI: 10.1016/j.jacc.2015.04.047.
- Haddad, Tarek, Adam Himes, and Michael Campbell (2014). “Fracture prediction of cardiac lead medical devices using Bayesian networks”. In: *Reliability engineering & system safety* 123, pp. 145–157.
- Haddad, Tarek, Adam Himes, Laura Thompson, Telba Irony, and Rajesh Nair (2017). “Incorporation of stochastic engineering models as prior information in Bayesian medical

- device trials”. In: *Journal of Biopharmaceutical Statistics* 00.00, pp. 1–15. ISSN: 15205711. DOI: 10.1080/10543406.2017.1300907.
- Henzinger, Thomas A., Peter W. Kopke, Anuj Puri, and Pravin Varaiya (Aug. 1998). “What’s Decidable about Hybrid Automata?” en. In: *Journal of Computer and System Sciences* 57.1, pp. 94–124. ISSN: 0022-0000. DOI: 10.1006/jcss.1998.1581.
- Hood, Richard (2015). *The EP Lab*. Accessed 10/20/2015.
- Huang, Samuel, Madeline Diep, Kuk Jin Jang, Elizabeth M. Cherry, Flavio H. Fenton, Rance Cleaveland, Mikael Lindvall, Rahul Mangharam, and Adam Porter (Nov. 2020). “Towards Automated Comprehension and Alignment of Cardiac Models at the System Invariant Level”. In: *CSBio ’20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*. CSBio2020. New York, NY, USA: Association for Computing Machinery, pp. 18–28. ISBN: 978-1-4503-8823-8. DOI: 10.1145/3429210.3429225.
- Ibrahim, Joseph G, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen (2015a). “The power prior: theory and applications”. In: *Statistics in medicine* 34.28, pp. 3724–3749.
- (Dec. 2015b). “The power prior: theory and applications”. en. In: *Statistics in Medicine* 34.28, pp. 3724–3749. ISSN: 02776715. DOI: 10.1002/sim.6728.
- Jang, Kuk Jin, Yash Vardhan Pant, Bo Zhang, James Weimer, and Rahul Mangharam (Apr. 2019). “Robustness evaluation of computer-aided clinical trials for medical devices”. In: *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*. ICCPS ’19. New York, NY, USA: Association for Computing Machinery, pp. 163–173. ISBN: 978-1-4503-6285-6. DOI: 10.1145/3302509.3311058.
- Jang, Kuk Jin, James Weimer, Houssam Abbas, Zhihao Jiang, Jackson Liang, Sanjay Dixit, and Rahul Mangharam (July 2018). “Computer Aided Clinical Trials for Implantable Cardiac Devices”. eng. In: *Annual International Conference of the IEEE Engineering*

- in *Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2018, pp. 1–4. ISSN: 2694-0604. DOI: 10.1109/EMBC.2018.8513284.
- Jenkins, Janice M. and Richard E. Jenkins (2003). “Arrhythmia database for algorithm testing: surface leads plus intracardiac leads for validation”. eng. In: *Journal of Electrocardiology* 36 Suppl, pp. 157–161. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2003.09.041.
- Jiang, Z., M. Pajic, and R. Mangharam (Jan. 2012). “Cyber-Physical Modeling of Implantable Cardiac Medical Devices”. In: *Proceedings of the IEEE* 100.1, pp. 122–137. ISSN: 1558-2256. DOI: 10.1109/JPROC.2011.2161241.
- Jiang, Z., M. Pajic, S. Moarref, R. Alur, and R. Mangharam (2012). “Modeling and Verification of a Dual Chamber Implantable Pacemaker”. en. In: *Tools and Algorithms for the Construction and Analysis of Systems*. Ed. by Cormac Flanagan and Barbara König. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 188–203. ISBN: 978-3-642-28756-5. DOI: 10.1007/978-3-642-28756-5_14.
- Jiang, Zhihao, Houssam Abbas, Kuk Jin Jang, Marco Beccani, Jackson Liang, Sanjay Dixit, and Rahul Mangharam (2016). “In-silico pre-clinical trials for implantable cardioverter defibrillators”. In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the. IEEE*, pp. 169–172.
- Josephson, M.E. (2008). *Clinical Cardiac Electrophysiology*. Lippincot Williams and Wilkins.
- Kalin, Ron and Marshall S. Stanton (Apr. 2005). “Current clinical issues for MRI scanning of pacemaker and defibrillator patients”. eng. In: *Pacing and clinical electrophysiology: PACE* 28.4, pp. 326–328. ISSN: 0147-8389. DOI: 10.1111/j.1540-8159.2005.50024.x.

- Lafferriere, Gerardo, George J. Pappas, and Shankar Sastry (2000). “O-Minimal Hybrid Systems”. English. In: *Mathematics of Control, Signals and Systems* 13.1, pp. 1–21. ISSN: 0932-4194. DOI: 10.1007/PL00009858.
- Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*.
- Lim, Jessica, Rosalind Walley, Jiacheng Yuan, Jeen Liu, Abhishek Dabral, Nicky Best, Andrew Grieve, Lisa Hampson, Josephine Wolfram, Phil Woodward, Florence Yong, Xiang Zhang, and Ed Bowen (Sept. 2018). “Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities”. en. In: *Therapeutic Innovation & Regulatory Science* 52.5. Publisher: SAGE Publications Inc, pp. 546–559. ISSN: 2168-4790. DOI: 10.1177/2168479018778282.
- Lin, Junjing, Margaret Gamalo-Siebers, and Ram Tiwari (2018). “Propensity score matched augmented controls in randomized clinical trials: A case study”. en. In: *Pharmaceutical Statistics* 17.5, pp. 629–647. ISSN: 1539-1612. DOI: 10.1002/pst.1879.
- Mahajan, Rajiv and Kapil Gupta (Aug. 2010). “Adaptive design clinical trials: Methodology, challenges and prospect”. In: *Indian Journal of Pharmacology* 42.4, pp. 201–207. ISSN: 0253-7613. DOI: 10.4103/0253-7613.68417.
- Makowiec, D., J. Wdowczyk, and Z. R. Struzik (2019). “Heart Rhythm Insights Into Structural Remodeling in Atrial Tissue: Timed Automata Approach”. English. In: *Frontiers in Physiology* 9. Publisher: Frontiers. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.01859.
- Mitchell, C. C. and D. G. Schaeffer (2003). “A Two-Current Model for the Dynamics of Cardiac Membrane”. In: *Bulletin of Mathematical Biology*. Vol. 65, pp. 767–793.
- Moss, Arthur J., Claudio Schuger, Christopher A. Beck, Mary W. Brown, David S. Cannon, James P. Daubert, N.A. Mark Estes, Henry Greenberg, W. Jackson Hall, David T. Huang, Josef Kautzner, Helmut Klein, Scott McNitt, Brian Olshansky, Morio Shoda, David Wilber, and Wojciech Zareba (Dec. 2012). “Reduction in Inappropriate Therapy

- and Mortality through ICD Programming”. en. In: *New England Journal of Medicine* 367.24, pp. 2275–2283. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1211107.
- Muntner, Paul, Paula T. Einhorn, William C. Cushman, et al. (Jan. 2019). “Blood Pressure Assessment in Adults in Clinical Practice and Clinic-Based Research: JACC Scientific Expert Panel”. eng. In: *Journal of the American College of Cardiology* 73.3, pp. 317–335. ISSN: 1558-3597. DOI: 10.1016/j.jacc.2018.10.069.
- Neuenschwander, Beat, Gorana Capkun-Niggli, Michael Branson, and David J. Spiegelhalter (Feb. 2010). “Summarizing historical information on controls in clinical trials”. eng. In: *Clinical Trials (London, England)* 7.1, pp. 5–18. ISSN: 1740-7753. DOI: 10.1177/1740774509356002.
- O’Hara, Thomas, László Virág, András Varró, and Yoram Rudy (May 2011). “Simulation of the undiseased human cardiac ventricular action potential: Model formulation and experimental validation”. In: *PLoS Comput. Biol.* 7.5, e1002061.
- O’Kelly, Matthew, Aman Sinha, Hongseok Namkoong, John Duchi, and Russ Tedrake (Dec. 2018). “Scalable end-to-end autonomous vehicle testing via rare-event simulation”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., pp. 9849–9860.
- Pappalardo, Francesco, Giulia Russo, Flora Musuamba Tshinanu, and Marco Viceconti (2018). “In silico clinical trials: concepts and early adoptions”. en. In: *Briefings in Bioinformatics*. DOI: 10.1093/bib/bby043.
- Pathmanathan, Pras and Richard Gray (2013). “Ensuring reliability of safety-critical clinical applications of computational cardiac models”. In: *Frontiers in Physiology* 4, p. 358. ISSN: 1664-042X. DOI: 10.3389/fphys.2013.00358.

- Pocock, Stuart J. (Mar. 1976). “The combination of randomized and historical controls in clinical trials”. English. In: *Journal of Chronic Diseases* 29.3. Publisher: Elsevier, pp. 175–188. ISSN: 0021-9681. DOI: 10.1016/0021-9681(76)90044-8.
- Rees, Johannes B. van, C. Jan Willem Borleffs, Mihály K. de Bie, Theo Stijnen, Lieselot van Erven, Jeroen J. Bax, and Martin J. Schalij (Feb. 2011). “Inappropriate Implantable Cardioverter-Defibrillator Shocks: Incidence, Predictors, and Impact on Mortality”. en. In: *Journal of the American College of Cardiology* 57.5, pp. 556–562. ISSN: 0735-1097. DOI: 10.1016/j.jacc.2010.06.059.
- Roos, Magorzata, Thiago G. Martins, Leonhard Held, and Hvard Rue (2015). “Sensitivity analysis for Bayesian hierarchical models”. In: *Bayesian Analysis* 10.2, pp. 321–349. ISSN: 19316690. DOI: 10.1214/14-BA909. arXiv: arXiv:1312.4797v1.
- Saxberg, B. E. H. and R. J. Cohen (1991). “Cellular Automata Models of Cardiac Conduction”. en. In: *Theory of Heart: Biomechanics, Biophysics, and Nonlinear Dynamics of Cardiac Function*. Ed. by Leon Glass, Peter Hunter, and Andrew McCulloch. Institute for Nonlinear Science. New York, NY: Springer, pp. 437–476. ISBN: 978-1-4612-3118-9. DOI: 10.1007/978-1-4612-3118-9_18.
- Schmidli, Heinz, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O’Hagan, David Spiegelhalter, and Beat Neuenschwander (Dec. 2014). “Robust meta-analytic-predictive priors in clinical trials with historical control information”. eng. In: *Biometrics* 70.4, pp. 1023–1032. ISSN: 1541-0420. DOI: 10.1111/biom.12242.
- Schoenfeld, David Alan, Dianne M Finkelstein, Eric Macklin, Neta Zach, David L Ennist, Albert A Taylor, and Nazem Atassi (Oct. 2019). “Design and analysis of a clinical trial using previous trials as historical control”. en. In: *Clinical Trials* 16.5. Publisher: SAGE Publications, pp. 531–538. ISSN: 1740-7745. DOI: 10.1177/1740774519858914.

- Simalatsar, Alena, Monia Guidi, Pierre Roduit, and Thierry Buclin (2019). “Methods for Personalised Delivery Rate Computation for IV Administered Anesthetic Propofol”. en. In: *Automated Reasoning for Systems Biology and Medicine*. Ed. by Pietro Liò and Paolo Zuliani. Computational Biology. Cham: Springer International Publishing, pp. 369–397. ISBN: 978-3-030-17297-8. DOI: 10.1007/978-3-030-17297-8_14.
- Spector, Peter S., Nicole Habel, Burton E. Sobel, and Jason H.T. Bates (2011). “Emergence of Complex Behavior: An Interactive Model of Cardiac Excitation Provides a Powerful Tool for Understanding Electric Propagation”. In: *Circulation: Arrhythmia and Electrophysiology* 4.4, pp. 586–591. DOI: 10.1161/CIRCEP.110.961524.
- Stroobandt, Roland X., S. Serge Barold, and Alfons F. Sinnaeve (Jan. 2009). *Implantable Cardioverter Defibrillators Step by Step*. Wiley-Blackwell. ISBN: 978-1-4051-8638-4.
- Ten Tusscher, K H W J and A V Panfilov (Sept. 2006). “Alternans and spiral breakup in a human ventricular tissue model”. In: *Am. J. Physiol. Heart Circ. Physiol.* 291.3, H1088–1100. ISSN: 0363-6135. DOI: 10.1152/ajpheart.00109.2006.
- Tivay, Ali, George C Kramer, and Jin-Oh Hahn (2021). “Collective Variational Inference for Personalized and Generative Physiological Modeling: A Case Study on Hemorrhage Resuscitation”. In: *IEEE Transactions on Biomedical Engineering*. Conference Name: IEEE Transactions on Biomedical Engineering, pp. 1–1. ISSN: 1558-2531. DOI: 10.1109/TBME.2021.3103141.
- Vambheim, Sara M and Magne Arve Flaten (July 2017). “A systematic review of sex differences in the placebo and the nocebo effect”. In: *Journal of Pain Research* 10, pp. 1831–1839. ISSN: 1178-7090. DOI: 10.2147/JPR.S134745.
- Vesper, Hubert W, Gary L Myers, and W Greg Miller (Sept. 2016). “Current practices and challenges in the standardization and harmonization of clinical laboratory tests¹²³”. In:

- The American Journal of Clinical Nutrition* 104.Suppl 3, 907S–912S. ISSN: 0002-9165. DOI: 10.3945/ajcn.115.110387.
- Viceconti, Marco, Adriano Henney, and Edwin Morley-Fletcher (May 2016). “In silico clinical trials: how computer simulation will transform the biomedical industry”. en. In: *International Journal of Clinical Trials* 3.2, p. 37. ISSN: 2349-3259, 2349-3240. DOI: 10.18203/2349-3259.ijct20161408.
- Vladimerou, Vladimeros, Pavithra Prabhakar, Mahesh Viswanathan, and Geir Dullerud (2008). “STORMED Hybrid Systems”. English. In: *Automata, Languages and Programming*. ISBN: 978-3-540-70582-6. DOI: 10.1007/978-3-540-70583-3_12.
- Wilkoff, Bruce L., Timothy Albert, Mariya Lazebnik, Sung-Min Park, Jonathan Edmonson, Ben Herberg, John Golnitz, Sandy Wixon, Joel Peltier, Hyun Yoon, Sarah Willey, and Yair Safriel (Dec. 2013). “Safe magnetic resonance imaging scanning of patients with cardiac rhythm devices: a role for computer modeling”. eng. In: *Heart Rhythm* 10.12, pp. 1815–1821. ISSN: 1556-3871. DOI: 10.1016/j.hrthm.2013.10.009.
- Wong, Chi Heem, Kien Wei Siah, and Andrew W Lo (Apr. 2019). “Estimation of clinical trial success rates and related parameters”. In: *Biostatistics* 20.2, pp. 273–286. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxx069.
- Yip, E., S. Andalam, P. S. Roop, A. Malik, M. Trew, W. Ai, and N. Patel (Mar. 2016). “Towards the Emulation of the Cardiac Conduction System for Pacemaker Testing”. In: *arXiv:1603.05315 [cs, q-bio]*. arXiv: 1603.05315.
- Zanker, Norbert, Diane Schuster, James Gilkerson, and Kenneth Stein (Sept. 2016). “Tachycardia detection in ICDs by Boston Scientific”. en. In: *Herzschrittmachertherapie + Elektrophysiologie* 27.3, pp. 186–192. ISSN: 1435-1544. DOI: 10.1007/s00399-016-0454-2.