# A Framework for Reasoning about Patient Safety of Emerging Computer-Based Medical Technologies

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Philip Kwame Danso Asare

May

2015

APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

_____

AUTHOR

The dissertation has been read and approved by the examining committee:

John Lach

_____

Advisor

John A. Stankovic

_____

Advisor

Joanne Bechta Dugan

_____

Gabriel Robins

_____

Stephen D. Patek

_____

_____

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

May

2015

# A Framework for Reasoning about Patient Safety of Emerging Computer-Based Medical Technologies

Philip Kwame Danso Asare

May 1, 2015

# Abstract

Like many industries today, the healthcare industry depends on computer-based technologies. From the digital thermometer to the magnetic resonance imaging (MRI) machine, one can find a variety of devices of different levels of complexity in any clinical environment. Mobile technologies are also driving many out-of-clinic solutions. The increasing complexity of medical technologies is improving both the diagnosis and treatment capabilities of the healthcare industry, resulting in improved patient outcomes. It is, however, also creating more opportunities for undesirable outcomes, with the primary concern being harm to the patients that these technologies are intended to help. This dissertation presents a framework for reasoning about this primary concern for computer-based medical technologies in light of this increase in complexity.

The core framework consists of a general model of patient safety based on a dynamical systems view of health and health management. It views the human body as a natural safety critical system and health as the body maintaining safe states. Doing so makes the goal of health management (where medical technologies are employed) aiding in this safety-critical function, and allows us to discuss safety of these technologies within the same framework used for discussing health. This provides a number of advantages. First, it makes the developments accessible to health practitioners. Second, it provides designers with a link between system design features and patient safety (viewed as health outcomes). Third, it provides regulators with a general framework for reasoning about the large number of instantiations of medical technologies. Most importantly, it allows all three stakeholders to reach a common understanding of patient safety for any medical technology. This makes the framework valid from a health management perspective. Casting health in safety terms makes it consistent with systems safety principles, while addressing the short-comings of existing techniques for dealing with health as a functional goal.

The ability of the framework to enabling reasoning about the complexity introduced by integration, autonomy, and mobility of emerging technologies is demonstrated by extending the core ides to one class of these technologies known as body sensor networks (BSNs). The result is a general set of hazards for BSNs based on a generic BSN model, and a proof-of concept simulation tool for BSNs embodying the features necessary for exploring issues related to safety. Realistic examples based on information from the literature are provided throughout to demonstrate the validity and applicability of the ideas.

# Preface

If you have ever been admitted to the hospital or watched a medical show on television, you will notice the wealth of devices and tests that are used to monitor the patient in order to make relevant medical decisions. The most familiar of these devices is the electrocardiogram: that device with the spikey waveforms that beeps to indicate the patient's heart is still beating and flatlines and goes into a monotone if the patient dies.

Today, the vision is that patients will be monitored outside of the clinical environment, using devices that can be implanted or worn, which will provide the necessary information for medical decision making. For example, a patient with a heart condition could be prescribed a number of devices that coordinate to provide his or her physician with information on the behavior of their cardiovascular system in relation to his or her daily activities. Naturally, because these new technologies are still medical devices, stakeholders are concerned about the safety of the patient.

This dissertation grew out of this concern. It started as collaboration with Paul Jones and Yi Zhang at the U.S. Food and Drug Administration (FDA), the federal agency tasked with approving medical devices. They were proactive in trying to develop ways to reason about the safety of these emerging technologies, which are unlike the systems they have traditionally had to approve, and to provide guidance to manufacturers on how to reason about and communicate the safety implications of these devices. I happened to be interested in modeling a version of these technologies called body sensor networks at the time, but was having a hard time making a value proposition on why anyone would need to model them. Paul and Yi provided me with a great one: reasoning about safety.

Why the focus on "reasoning"? The search I just ran on oxforddictionaries.com provided a definition of "reasoning" as "the action of thinking about something in a logical sensible way." When we design a medical technology, we have to make some decision on whether it will be safe for patients. Certainly we would like this decision-making process to be "sensible". We would also like it to be "logical" or at least transparent in a way that someone else can follow and question our rationale and assumptions. Medical technologies are complex, and understanding how they can become unsafe is riddled with subtleties.

The first question I asked myself is "what exactly do we mean when we say a medical device is safe?" If I could find an answer to this, then I could apply the same basic thinking, or update it for these emerging technologies. The answers I got, when I did get any, seemed quite unsatisfactory, so I decided to supply my own. The framework presented in this dissertation is based on my current thoughts on what this answer is. It is based on what I have learned from how people think (reason) about safety of technologies and why and how we go about managing health using various medical technologies. In the process, I found that one of the difficulties in answering this question lies in the ambiguity of health as a concept. What does is it mean to be healthy or unhealthy? When should we intervene to try to improve health? My contention is that despite the variability in the answers to these questions, one should be able to express their rationale for a particular answer precisely.

The dissertation focuses on ways for expressing this rationale precisely. If we can express what it means to be health or unhealthy, and hence why we should build a medical technology (and in effect how to measure its efficacy), maybe we can use the same approach to express when we think a medical device is safe (healthy) for a patient population. Since this topic is of interest to a wide range of people, I tried to keep the writing for a broad audience. I hope that you find something insightful and useful whether you are a designer interested in developing a particular medical technology, a regulator tasked with evaluating one, a health practitioner trying to understand the claims made about one, or just generally interested in the topic.

—- Philip Kwame Danso Asare,
Charlottesville, Virginia, 2015

# Acknowledgments

I'd first like to thank God for keeping me going throughout the process and for putting me in situation to meet the wonderful people I am about to thank next, without whom many of the ideas here would have never seen the light of day.

First are my advisors, John Lach and Jack Stankovic, who are my academic parents. I am grateful for their patience through the ups and downs, especially when I would throw the equivalent of academic adolescent tantrums. Their guidance has helped me grow from an academic adolescent into an academic adult. John showed me that it is okay to be non-traditional, and that if you weren't annoying somebody, then you probably weren't doing something worthwhile. Jack indulged me when I struggled to communicate an idea that was not yet fully formed ("crisp" as he would call it), and asked enough questions to help me move on to the next step. He also reminded me that no matter how good I thought an idea was, there was always some way to make it better. Both were supportive in my many excursions that enriched my PhD experience and provided the external inputs I needed to refine ideas and keep motivated.

Second is my dissertation committee. Joanne Dugan also spent quite some time listening to ideas and providing feedback, and always seemed excited at my presence even if it was the third time I had been in her office that day. She also helped me in running the independent study course the made the development of BodySim possible. Gabe Robins asked insightful questions and made remarks at committee meetings that helped me understand what I was trying to do. Steve Patek was a great source of external validation. He introduced me to the folks at the Center for Diabetes Technologies (CDT) and TypeZero Technologies who provided me with many resources and from whom I learned a lot about safety issues in the artificial pancreas and general issues in diabetes management.

The folks at the CDT welcomed me to their meetings, gave me an audience when I had an idea to present, and provided valuable feedback. Andy Ortiz, in particular, was instrumental in getting the simulations for the case studies done, and Marc Breton provided me with data from previous human subject trials. Many of the examples throughout the dissertation would not exist without their help.

Chad Rogers at TypeZero assured me that the "reasoning" problem was where the issue was and that he could get talented team of engineers to build a system so long as they knew what it meant for that system to be safe.

Next is the FDA. Paul Jones and Yi Zhang were instrumental in getting me the fellowship which helped with starting with this work. They trusted me to take a stab at the problem and gave me space to do some good thinking on the issue when I was in-residence there. They also pointed me to valuable resources. Had Paul Jones not encouraged me to go to the AAMI/FDA workshop on medical device interoperability, I would not have come across Nancy Leveson's work and discovered systems safety. They introduced me to Sandy Weininger who pointed me to Webster's work on medical instrumentation, which helped me think through the body sensor network problem.

Members of the INERTIA Lab and the Wireless Sensor Network Lab past and present provided support and motivation. Some I would like to mention are Italo Armenti and Juliana Su, with whom I worked on quality of information which shaped some of my thinking on body sensor network hazards. Shanshan Chen was always happy to chat about ideas. Jeff Brantley helped me settle into the group. Jiaqi Gong gave me the needed distraction from abstract ideas by providing concrete problems to discuss and work on. The newer students (Luis, Josh, David, Ben, and Sriram) have helped with various aspects of BodySim. Rob Dickerson and visiting scholar Xianyue Wu helped me establish the feasibility of BodySim. Rob created the first human subject model based on data we collected on Xianyue. Without them, it would have just been an idea. Victor Sobral, another visiting scholar has been instrumental in continued development.

Speaking of BodySim, I must mention Scott Tepsuporn, who stayed on after the independent study to bring it from prototype for demo at a conference to a functioning proof-of-concept version with many interesting features.

Interactions with David Stone and Marie Csete who he introduced me to helped me solidify some of my thinking on viewing humans as dynamical systems.

Lastly, I must thank family and friends for their continued support, even when it seems like I have not had enough of being in school. The meaningful discussions we had about my work continues to convince me that good input can come from many sources. Gloria Aghaulor, my soon-to-be wife deserves special mention for enduring various rants and excited presentations of ideas, especially on days when neither were warranted.

*Me da mo nia ase.*

# Contents

# Chapter 1

# Introduction

> Solving a problem simply means representing it so as to make the solution transparent
>
> Herbert A. Simon

Like many industries today, the healthcare industry depends on computer-based technologies. In the clinic, for example, digital thermometers help take temperature, patient monitors collect vital signs and alert nurses when values go out of the normal range, magnetic resonance imaging machines provide information on internal anatomy, and radiation therapy machines irradiate specific parts of the body to treat various undesired growths. Recently, advances in mobile and networking technologies are enabling patients and healthcare providers to manage care away from the clinic. The ever-increasing complexity of medical technologies* is improving diagnostic and therapeutic capabilities, resulting in improved patient outcomes. However, there is also an increased potential for harm to the patients that these technologies are intended to help. This dissertation presents a framework for reasoning about this primary concern for these computer-based medical technologies in light of this increase in complexity.

## Patient Safety of Medical Technologies

The potential for harm of any system falls within the realm what is generally called "Systems Safety Engineering." According to a primer by the New England Chapter of the

*"Medical technologies" refers both to individual devices or networks devices.

System Safety Society [104], quoting the Military Standard MIL-STD-882, this is "the application of engineering and management principles, criteria, and techniques to achieve acceptable mishap risk, within the constraints of operational effectiveness and suitability, time, and cost, throughout all phases of the system life cycle."

This dissertation is primarily concerned with the idea of acceptable mishap risk. Mishaps are generally things we do not want to happen. Risk has to do with the potential for and severity of occurrence of these mishaps. The word "acceptable" implies that there has to be some consensus by different stakeholders on what are the mishaps and what level of risk is deemed acceptable. Below, these ideas are put in the context of medical technologies.

Medical technologies are typically designed for a particular patient population. General Electric's Dash 4000 patient monitor [38], for example, is designed to be used for a general patient population. Because of the complexity of the human body and how its characteristics vary from person to person, in the use of a particular medical technology, some set of patients will almost invariably be harmed. For example, radiation therapy may have no side effects in some patients, while it could produce secondary cancers in others [67].

We are therefore generally concerned with how much of this harm is tolerable, in terms of how many patients get harmed and how severe the harm is. A system that results in a tolerable amount of harm (acceptable mishap risk) is considered safe for patients. Hence, in general, radiation therapy systems are considered to be safe for patients because we are willing to tolerate the number of people who experience severe effects like secondary cancers.

The above implies that patient safety is not an absolute concept. Because what is considered safe and unsafe for patients depends on this tolerance level for harm (which in turn depends on a number of factors including the ambiguous concept of health), when our tolerance changes so does the patient safety of the technology. One of the things this dissertation shows (in chapters 3 and 4) is that despite this relativity of what is considered tolerable harm for patients, and the ambiguity of some of the ideas this tolerance depends on, we can still have precise and meaningful discussions when reasoning about the patient safety of medical technologies.

## Stakeholder Reasoning Needs

Three main stakeholders are involved when it comes to the safety of medical technologies for patients: the technology manufacturers who design and sell medical technologies

intended to help health practitioners[†] carry out their health management duties more effectively; the health practitioners who work with these technologies to help patients; and the regulatory agency who must protect patients by ensuring that only safe technologies are approved for sale. In the U.S., the regulatory agency is the Food and Drug Administration (FDA)[‡].

Ideally, all three stakeholders would have a common understanding of what "patient safety" (the acceptable mishap risks) for a particular technology means. Without this, there would be confusion on whether particular outcomes for patients are undesirable or not, and on how to address a particular technology of concern. Ideally, patient safety would be defined in way that is relatable to contexts in which these technologies are used and to patient outcomes, making it accessible to health practitioners. Using a radiation therapy machine as a hypothetical example, we would say it is safe because when used as instructed on cancer patients only a small percentage (say 1 in 100,000) experience secondary cancers, and of this number only 1 in 10,000 are malignant.

To be accessible to manufacturers, this definition of patient safety in terms of patient outcomes must be linked to design features of the particular technology. This link should enable manufacturers to address patient safety in a way that fits in their development process. For example, if it turns out that the outcomes (risk of secondary cancers) for our hypothetical radiation therapy machine are linked to the accuracy of the focus of the radiation beams, then manufacturers can use this link between patient outcomes (secondary cancer risk) and design features (beam accuracy) in their design considerations.

If the manufacturer is interested in making improvements to the machine design, they would make sure that either the beam accuracy remains at levels required for the outcomes we currently tolerate or change in a way that result in better outcomes. In addition, if a manufacturer wants their machine to also be used for non-cancer patients to treat other growths, they would have to understand what the tolerable outcome for this population is and how these relate to the accuracy of beam focusing. If the tolerance is different for cancer and non-cancer patients, this would affect the design requirements.

To aid the FDA in their regulatory function, the definition of patient safety and the resulting framework for reasoning about patient safety must provide a means for examining

[†]I use the term "health practitioners" to refer to those in the medical field and other fields related to human health, including researchers in these fields.

[‡]According to the FDA website, "FDA is responsible for protecting the public health by assuring the safety, efficacy and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation"[110].

in a consistent manner technologies from different manufacturers with similar health management goals. For example, if ten different companies make radiation therapy machines but use different techniques to focusing the beams, the FDA would be more concerned with the accuracy of that focus since that is what is linked to the patient outcomes, and not necessarily the details of the particular techniques.

A general way of defining patient safety that satisfies the above criteria provides all three stakeholders with a repeatable mechanism for collaboratively arriving at specific patient safety definitions for any medical technology of concern. This approach to defining patient safety is important for both current and emerging technologies. Typically, the expectation is that the FDA bears the sole responsibility of defining what is safe. As proactive as the FDA tries to be, it cannot anticipate all new technologies. In addition, it is impossible for the FDA or any other regulatory agency to understand the details of every existing and emerging medical technology, especially as the variety and complexity of these technologies increase. Letting the FDA bear the sole burden of defining what is safe is therefore impractical.

Relating health management contexts and patient outcomes to technology design features allows any of the three stakeholders to examine meaningfully how changes to a design affect patient outcomes or how changes in patient outcome expectations affect designs, and hence patient safety. The ability to do this is important for striking a balance between feasibility of designs and patient safety, a process that requires input from all three stakeholders. For our example radiation therapy machine, all three could have a discussion on what amount of potential occurrences of secondary cancers is tolerable knowing that it affects the accuracy of beam focusing required and hence the feasibility of the design (both in terms of cost and what is actually possible given the current state of engineering).

For emerging technologies or substantial revisions to existing technologies, this approach allows the manufacturer to develop a preliminary definition of safety which can be refined with input from the other stakeholders. This prevents these manufacturers from waiting on the FDA to make sense of their technology before beginning to think about its patient safety implications. The earlier safety considerations can be incorporated into the design process, the better [104].

## Complexities of Emerging Medical Technologies

The increase in complexity of emerging medical technology that motivates this dissertation is fueled by three recent trends. First is the integration of medical devices that have previ-

ously operated independent of each other into systems with newer capabilities for diagnosis and treatment. Second is the increase in autonomy that both these integrated systems and individual devices are being given, where some decisions that were previously being made by health practitioners are now being made by devices. Third is the development of mobile systems to be used away from the clinical environment, typically by the patient and without direct supervision of a health practitioner or immediate availability of technical personnel.

Previously, in the absence of these trends, much of the safety burden lay with the health practitioner who used these devices, making reasoning about patient safety of the devices much simpler. These trends, however, redistribute some of the safety burden to the devices, creating the need to reexamine patient safety accounting for the increased complexity, and to handle some of the newer issues introduced.

A category of emerging medical technologies that combines all three trends is what is know as body area networks (BANs). An example is the artificial pancreas [29], a system designed to help Type I diabetics manage their blood glucose levels in real time. Type I diabetes is an auto-immune disease that prevents the body form producing insulin, a hormone needed keep blood glucose levels from becoming abnormally high [73]. Current versions of the artificial pancreas typically consist of three physically-separate wearable devices: a device for monitoring the blood glucose, a pump for infusing insulin, and controller device which uses the information it receives from the monitoring device, as well as other input from the patient, to decide on and instruct the pump on how much insulin to infuse.

An interesting issue arises in the case of systems like the artificial pancreas. Although not currently on the market, the prototypical artificial pancreas has its three devices developed by three different manufacturers. In such a case, each manufacturer must have their device approved by the FDA independent of the other. It makes sense to talk about the patient safety of the artificial pancreas as a whole, but what does it mean to say that any of the devices on their own is safe for patient? And, how do we define this idea of patient safety of the individual devices that make up the system such that if the manufacturer intends that their device would be used in an artificial pancreas, then we do not require that the manufacturer has to test their device with every version of the other devices available on the market? (Each device can be used for other purposes besides being part of the artificial pancreas.) More importantly, how do we make sure there is no confusion between the different manufacturers, the FDA, and health practitioners on meaning of patient safety for a device? This problem affects medical devices that have the ability to be integrated into a larger system.

## The Systems Safety Gap

The obvious question to ask is that if there is a whole field of systems safety, what is missing, and why the need for the work in this dissertation? Above, we already mapped some general systems safety ideas to the medical context. The key lies in the difference between technologies like those used in the medical context we are concerned with and other safety-critical systems.

It turns out that we can put safety-critical systems into two general categories: those where human health as a safety issue, and to a large extent safety in general, is considered as a constraint; and those where human health as a safety issue is the functional goal.

Examples of the first kind of system are automotives, chemical plants, and nuclear plants. The functional goal of a vehicle is to transport people and goods from one point to another, and the goal of a nuclear power plant is to generate electricity for industrial and home consumption. When a person falls sick, our first inclination is not to go build a nuclear power plant. However, whenever a nuclear power plant is built (for other reasons), we are definitely be concerned with the health risks and other potential for harm to people. One way to reduce risk may be to not build the power plant at all, and the risk is introduced because we built the power plant.

Medical technologies fall into the second kind of system, and so do water treatment plants. For these systems, their functional goals are measured with respect to the outcome for human health and well-being. When a person falls sick, we seek solutions where medical technologies play a role, and in some cases preventative solutions like water treatment plants. This is why the previous section stresses the importance of putting patient safety in patient outcome terms.

The main issue is that systems safety grew out of the first kind of system, and hence its current treatment of human health as a safety issue is very limited. There is limited precision with what is described as harm to human health since that is not the primary focus of the system design, and hence when applied to medical technologies, this reduces the ability to effectively address risks directly related to patient outcomes.

There is another problem with systems safety when it comes to dealing with human health. In the first kind of system, the parts of the system that are considered in detail and that we are trying to make safer are technologies of our own creation, where as in the second kind, what we are concerned with are natural processes that we did not design. For a car, for example, we are concerned with its safety implications for the driver and others.

Since we designed the car, we can control the design process to add features that reduce its potential for harm and the severity of harm.

In the case of medical technologies, one can think of the human body as a natural safety-critical system that tries to keep itself safe (healthy) and the goal of medical technologies to aid in this safety critical function. This means that the medical technologies must work well together with the existing mechanisms of the human body and hence these mechanism must be understood and incorporated into the safety discussion. This case is analogous to being given a car without any knowledge of its design documents and being asked to add safety features without this knowledge. Since our understanding of the workings of the human body is still incomplete, it is more difficult to understand how it might potentially interact with new technologies.

There are a number of efforts in the medical technology community aimed at increasing capabilities for assuring patient safety. Many of these are based on applying traditional approaches to systems safety from other fields to the medical technology context, and hence make weak and vague links to actual health management contexts. They are also based on more mature medical technologies for which mishaps are well known. In recent years, for example, there has been a call for model-based approaches to medical technology design [2, 57], and there is a general shift in the research community and by the FDA towards this approaches for addressing patient safety of medical technologies [47]. There is also interest in dealing with patient safety in cases where medical devices are integrated and given more autonomy [5, 78]. Some work has also been done for emerging technologies like continuous glucose monitors [52].

It is important to note, however, that the systems safety techniques that have been developed in the context of the first kind of system, and applied to existing medical technologies, are not completely useless for addressing emerging medical technologies. What is needed is reorienting them towards the view necessary to address patient safety for emerging medical technologies. This is exactly what this dissertation does: it provides the missing pieces to complement the existing systems safety approaches. The thesis is that from systems safety principles and careful abstraction of the health management context, one can derive a framework for reasoning about patient safety that complements existing techniques. The argument is that the framework is valid from a health management perspective (which is one of the reasoning needs for the stakeholders of concern) because it uses a careful abstraction of this context, and that it is consistent with systems safety principles because it was derived from these principles.

# Contributions of this Dissertation

The main contributions of this dissertation is in developing a framework (a model) for reasoning about patient safety that is both valid from a health management perspective and consistent with systems safety. The framework is based on dynamical systems ideas. This is because both systems safety and health management can be viewed in terms of dynamical systems, which allows us to ensure both validity and consistency. The aim of the framework is to provide as general a treatment of the topic as possible.

This framework does *not* constitute a formula for getting a product approved by the FDA. The intention is not to develop such a formula, and I find the idea of the existence of such a formula problematic. Below are brief descriptions of the pieces of this dissertation that make up the framework and help demonstrate the thesis.

## Core Ideas

The developments below focus on the core ideas behind the framework that are widely applicable to different emerging medical technologies. They are based on the realization that both health management and systems safety can be described using dynamical systems ideas, and by viewing the human body as a natural safety critical system, patient safety of medical technologies can be reasoned about using the same ideas used to discuss health (which is viewed as the body's attempt to maintain 'safe' function).

### Extracting Systems Safety Principles

Chapter 2 reviews systems safety, its goals, and some of the general approaches that have been developed to aid in achieving the goals. Its main focus is to extract the principles which guide the development of the other ideas in the dissertation. It also is the beginnings of providing a framework for examining different systems safety paradigms and approaches. In particular it shows that a dynamical systems view of safety is a valid general case of many of the approaches of systems safety, and shows how particular paradigms and approaches can be viewed as special cases of this more general view.

### A Model of Health Management

The primary goal of medical technologies is to aid in health management, and all discussions of these technologies must be related to this goal—some, especially manufacturers

of emerging technologies, sometimes lose sight of this point. The framework development begins in chapter 3 by developing a dynamical systems model of health management viewing the human as a 'natural' safety critical system and health as the ability to sustain this safety-critical function. This view ensures consistency with systems safety principles.

Though viewing humans and their health through a dynamical systems lens is not a new idea, a new notion that comes out of this development is the general structure and interpretation of health metrics in dynamical systems terms. Previous dynamical systems modeling of health management (which aid the validity argument for the framework developed in this dissertation) have been focused on explaining disease mechanisms or developing treatment strategies, but none, as far as I am aware, has looked specifically in general at what health metrics really are.

Health metrics are important for determining when to intervene and how well an intervention does. For emerging technologies, the ability to develop appropriate metrics (when they do not exist) is important. Since they are the measure of well the body manages to keep itself healthy (and hence 'safe'), health metrics also form the core piece in our model of patient safety in chapter 4. To demonstrate the validity and generality of the model from the perspective of health management, the ideas in the model are mapped to ideas in health management using a variety of examples from the literature and other credible sources.

### A Model of Patient Safety

Chapter 4 builds on the dynamical systems model of health management to show how mishaps and risks arise in the context of health management where medical technologies are employed. It reveals the role of variability in health management mishaps and develops the model to make sure variability is explicitly accounted for and hence plays a central role in the definition of patient safety (acceptable mishap risk).

Using two illustrative case studies related to different glycemia management contexts (one for the intensive care units (ICU), and the other for the artificial pancreas), it shows the utility of the model for defining patient safety criteria, assessing patient safety of specific medical technologies, developing requirements for a safety-guided design process, discussing safety criteria in light of design feasibility issues, exploring the impact of assumptions and rationale made in defining criteria and performing assessments, and structuring safety arguments.

The case studies focus on glycemia management because I had ready access to models, data, and resources locally (through the University fo Virginia Center for Diabetes Tech-

nologies) to make sure the case studies, though illustrative, were meaningful. Since the core ideas (developed in chapter 3) on which this model rests were shown to be applicable to cases other than glycemia management, one can argue that the patient safety model applies more generally as well.

## Implications and Applications

These developments focus on using the ideas described above to reason about patient safety for a particular class of emerging technologies, in this case body sensor networks, by looking at the implications for two different aspects of the design process. The first is the implication for patient safety analysis based on hazard and causal factor identification as is traditionally done in systems safety engineering. This shows how the framework complements existing techniques. The second is the implications for design tools focusing on features necessary for exploring the effect of interactions between BSNs and the patient on BSN behavior, an issue identified by previous developments as important for patient safety.

### Implications for Safety Analysis of Body Sensor Networks

As mentioned previously, manufacturers are often faced with the case of reasoning about the safety of a technology that may be part of a larger system. BSNs are such a technology and their patient safety is more subtle to reason about because of the indirect effect they have on patient outcomes. One of the implications of the model developed in chapter 4 is that by defining appropriate interfaces between the technology of concern and other subsystems in the health management system, one can reason about safety that particular technology.

Chapter 5 develops a generic BSN model and uses these to identify a general set of hazards for BSNs, the first such model and set of hazards as far as I am aware. The model and hazards reveal how the interface between the BSN and hazards must be characterized in order to undertake a safety-guided design process. In addition, it provides a discussion of some of causal factors for the hazards and develops a (non-exhaustive) set of points to consider in order to inform patient safety analysis of BSNs. As with other developments, examples are provided of how these ideas map to realistic systems to demonstrate their validity and utility.

### Implications for Body Sensor Network Design Tools

Since the human body is non-uniform in space, where (and how) a BSN component is worn (or implanted) matters. In addition, the physical characteristics and activities of the patient can heavily affect BSN behavior. Design tools must therefore allow a designer to explore these factors and interactions and their effect on BSN behavior.

Chapter 6 looks at these implications for BSN design tools. In particular, it focuses on tools for simulation-based explorations and virtual prototyping. It describe the characteristics needed in design tools in order to enable such explorations, and presents an open-source simulation framework (still under development) that serves as example instantiation of a tool with these characteristics. This framework is called BodySim, and its current version supports inertial sensing explorations. It also tracks variables that can be used to develop wireless communication models for BSNs.

BodySim provides realistic virtual humans that a user can place sensors on and run simulations similar to the way a real human subject experiment would be run. This virtual human subject experimentation platform eliminates the overhead of real human subject experiments, while providing advantages like repeatability in subject behavior and a level of control over experiment variables that is not possible in real human subject experiments.

It also provides advantages over existing (more issue-specific) models because of its multi-domain simulation capabilities and its ability to serve a framework for integrating different models to provide more meaningful evaluations of system behavior. This allows more intuitive joint simulation of sensing, processing, and communication in a BSN while accounting for the complex dynamic environment in which these must take place.

## Perspective

In 1940, Claude E. Shannon produced a masters thesis at the Massachusetts Institute of Technology titled “A Symbolic Analysis of Relay and Switching Circuits,” in which he developed a mathematical method for analyzing the properties of switching circuits and for designing circuits to exhibit specific properties [93]. Switching circuits and relays already existed and were being used in telephone exchanges and other equipment. Their design was, however, more of an art. By providing a formal model of these circuits, Shannon paved the way for more systematic design techniques (what we know today as Boolean Algebra for Digital Logic Design), and we have him to thank for the increase in complexity of circuits that are responsible for today’s impressive digital technologies.

In 1948, Shannon, then at Bell Labs, published an article titled "A Mathematical Theory of Communication" for the The Bell System Technical Journal [94]. Various impressive communication technologies already existed, and communications engineers were busy trying to improve them. These included the telegraph (including the transatlantic systems), telephone, radio, and television. In his article, Shannon summed up the fundamental goal of all communication systems: "reproducing at one point either exactly or approximately a message selected at another point." He then developed a theory around this idea by introducing the idea of a 'bit' (a term he attributes to John Tukey) as a unit of information to be communicated, describing the now-common schematic model of communication systems, and dealing with the issue of encoding messages and reliably communicating them over noisy channels. The formal concepts of information and communication developed in Shannon's theory provided communication engineers with invaluable tools for understanding and designing communication systems, and gave birth to the field of information theory that has influenced many developments. One could argue that Shannon's two theories are single-handedly responsible for the information technology revolution.

It is important to note that in either case, Shannon did not invent all the formalisms that he used. The Boolean Algebra he used for his circuit work, for example, was developed by George Boole [13]. Shannon's main contributions, in both cases, were in his insights and in articulating the implications of these insights in terms of design and analysis approaches for engineered systems. In the circuits case, he realized the analogy between two-state behavior of switching circuits and the two-valued logic abstraction of Boolean Algebra and the advantages it provides for describing the analysis and synthesis of such circuits [13, 39].

This phenomenon of formalizing an already-existing concept, system, or activity is a recurring theme in the history of technological developments. The class of systems known as "governors" (the well-known being the Watts governor for controlling the speed of steam engines) were developed without any rigorous mathematical theory. However, it was difficult to explain how these systems could enter a state of what is now known as 'instability" where they would oscillate instead of maintaining a steady state, hindering the ability to devise sound solutions to the problem. Work by James Clerk Maxwell [66] and others kicked off the field of control theory which helped solve the problem and enabled the design of the complex control systems that made the aircraft possible and keeps many systems like cars and chemical plants running smoothly.

In some sense, history is on our side when it comes to formalizing an already-existing concept. Indeed, Lee and Varaiya make the claim in their text on signals and systems [55] that "one way to get a deeper understanding of a subject is to formalize it, to develop math-

ematical models." This work does not claim to be "the theory of patient safety of medical technologies", but these historical examples are good models to follow, and have serve as inspirations for this work. In the spirit of Shannon and others, the hope is that in some way, however small, the work presented in this dissertation helps fuel advancements in our ability address patient safety of medical technologies, and paves the way for improvements in patient outcomes balanced with the potential for harm.

# Part I

# Core Ideas

# Chapter 2

# The Many Faces of Systems Safety

> What's in a name? That which we call a rose by any other name would smell as sweet.
>
> William Shakespeare

## Chapter Overview

This chapter is a brief review of some of the ways of reasoning about safety of systems. Its aim is not to be a comprehensive review of safety analysis techniques, but to present the paradigms that guide these techniques in order to put both existing techniques and the work in this dissertation in some context.

The basic premise in this chapter is that systems safety paradigms share a common dynamical systems view of the operation of a system, but differ in their conceptions of how emergent behavior, especially as it pertains to mishaps, arise. In addition all techniques share the common goal of identifying (their view of) causes of mishaps and controlling these causes to reduce the frequency and severity (risk) of these mishaps to an acceptable level.

## 2.1 Introduction

The goal of systems safety engineering is to be an integral part of the full life cycle of the system. An illustration of a system life cycle is shown in Figure 2.1. At the earliest part

of the cycle is the conception of the system, where a specific problem or need is identified. At conception, we envision the ideal operation of the system with the aim of producing an implementation which when deployed solves the particular problem or meets the particular needs. Most implementations will stem for a particular design, and designs would be based on requirements developed from the ideas and needs provided at the conception stage. At the end of life (useful operation) of an implementation of a particular system concept, the system is retired.

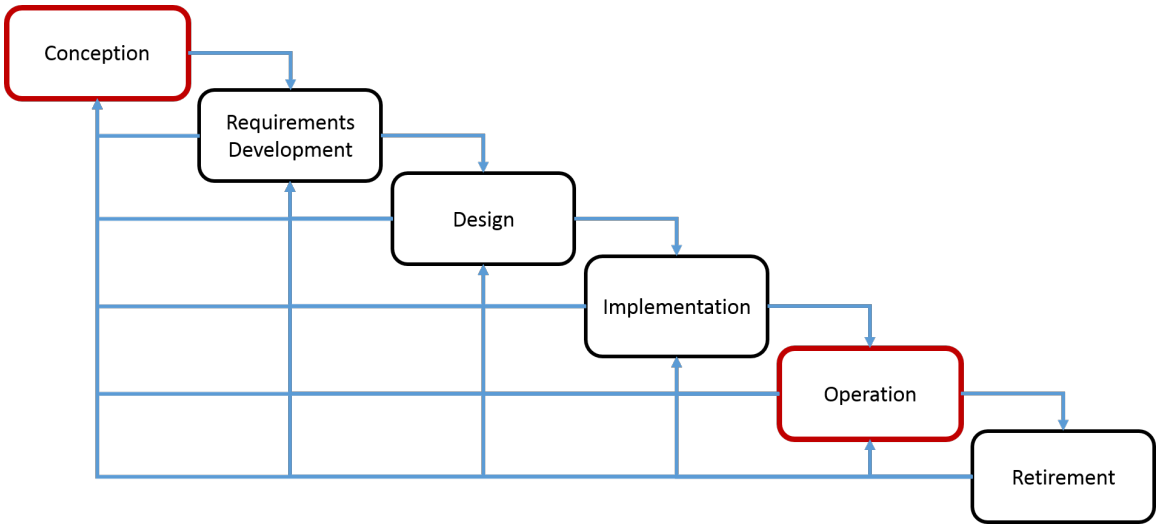



Figure 2.1: Illustration of system life cycle based on a modified version of the popular Waterfall Model.

The main aim of systems safety engineering is to ensure that from operation (when the system comes to physical existence) to retirement, the systems existence results in acceptable mishap risk. Ideally, safety considerations would start at conception. In addition, a large part of the systems operation (including operating procedures for humans who play an integral role in the system) is controlled by the design and implementation, which are in turn (ideally) controlled by the requirements. In the system safety engineer's ideal world, the requirements would ensure safety, and safety would be designed into the systems operations. If the operation does end up in producing mishaps, analysis of the mishap would provide valuable information on how to avoid mishaps for future versions of a similar system.

Much of systems safety is focused on how safety can be introduced at the earlier stages to inform design. Reasoning about safety requires making inferences about how

the conceived system might behave and whether those behaviors could potentially result in mishaps. This reasoning process is involves analysis of different versions of designs (and possibly implementations) of the system, and using the results from these analyses to inform design changes. Below is a review of some systems safety paradigms and techniques (with some historic context). In particular, it shows how the techniques share the common view point of dynamical systems, but differ in which systems components they focus on and their assumptions on how overall system behaviors, especially mishaps, arise.

## 2.2 Dynamical Systems

To avoid any confusion, it is important to briefly describe the particular view of dynamical systems used.

A dynamical system, for our purposes, is simply an abstraction for a system that evolves over time. It has states ($x$), a set of variables whose current values can (in theory) be used to predict future evolution of the system. It has parameters ($\lambda$), which govern its evolution. As part of its evolution, it may react inputs ($u$) to produce observable outputs ($y$). The values of any of these variables over time is call a trajectory. Visually, the structure of a dynamical system is represented by a box with arrows indicating the input and outputs as shown in figure 2.2.

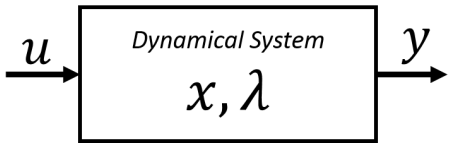



Figure 2.2: A visual representation of a generic dynamical system.

Below, we review a number of ideas useful for our purposes.

### 2.2.1 Trajectories

A trajectory is simply the description of the values of any of the variables (inputs, outputs, states, or parameters) as a function of time. Figure 2.3 shows three different trajectories related to an example model of a simple bank account whose basic structure is shown in figure 2.4.

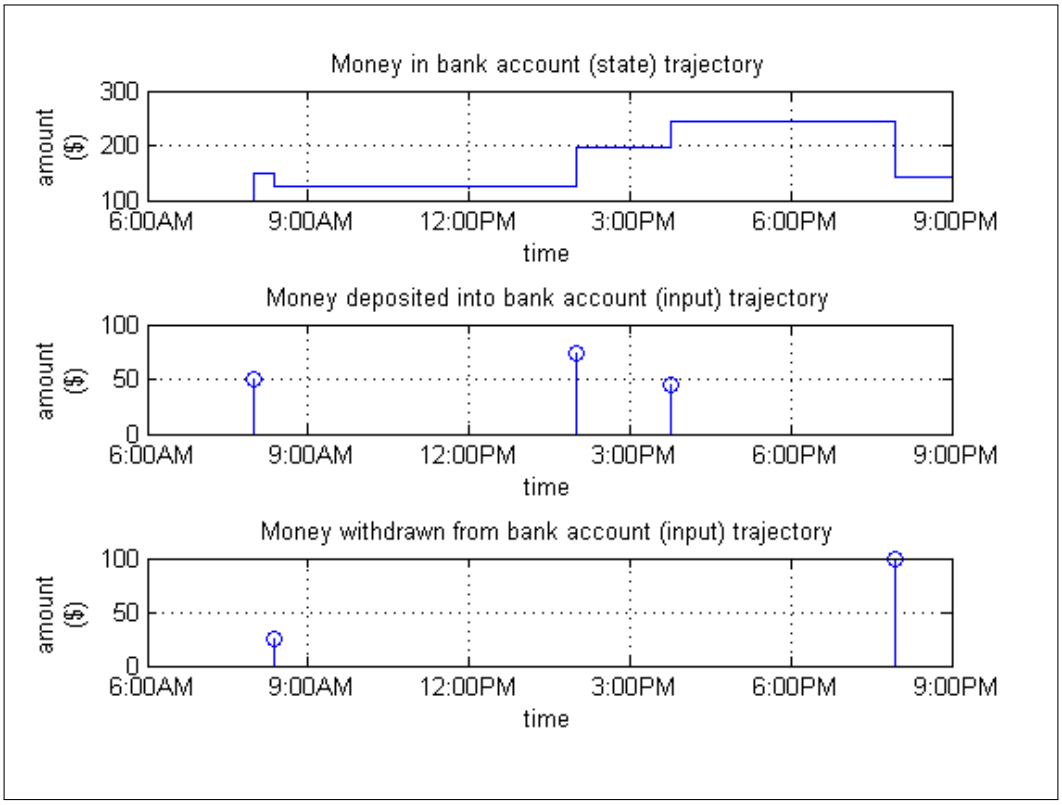



Figure 2.3: Example trajectories of bank account inputs and state

The first trajectory is the money in the account, which in this example is the state. The second and third trajectories represent deposits and withdrawals respectively. Notice that withdrawals are considered inputs, even though this seems counter intuitive from what happens in reality. From the perspective of modeling the money in the account, however, this makes sense since one can think of them as negative deposits.

### 2.2.2 Dynamics

The dynamics are concerned with the ways in which the system can evolve over time and the relationships between the various variables.

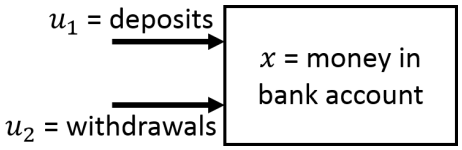



Figure 2.4: Structure of simple bank account model

### Deterministic Behavior

The simplest way to explain deterministic behavior of dynamical system is that given the same initial conditions and input trajectories, it will always evolve the same way. Using the example account in Figures 2.3 and 2.4, if we start at $100 at 6am, and have deposits and withdrawals follow the same pattern, the money in the account over time will always follow the same pattern as shown in Figure 2.3.

### Non-Deterministic Behavior

For a non-deterministic system, given the same initial conditions, and input trajectories, the evolution of the system may be different each time. One can think of a non-deterministic system as having some degree of 'free will' or randomness. Continuing with bank account examples, if we assume that we have an interest earning account, but the bank can arbitrarily change the interest on the account at any point in time, then given the same initial amount in an account, and the same deposit and withdrawal patterns, the way the amount in the account varies will be different depending on how the bank decides to adjust interest, which is a choice we know nothing about before hand.

### Modal Behavior

The behavior of a dynamical system can include more 'symbolic' and less quantitative. In the bank account example, an account may be designed such that so long as certain conditions exist, for example, if the balance on the account is above a certain minimum, we may not be assessed fees. In this case, we can think of the bank account as having two modes: a fee-based mode and a fee-free mode.

The dynamics (how the account behaves) in each mode is slightly different. If the account is not interest-earning, for simplicity, then in the fee-free mode, its behavior depends on the initial amount in the account and the input trajectory of withdrawals and de-

posits. Once conditions change and the balance goes below the minimum, then the account switches to a fee-based mode where in addition to responding to withdrawals and deposits, the account also reduces the balance by the fee amount periodically. Mode changes can deterministic or non-deterministic. The only requirement is that the possible modes are described as part of the description of the system, even if changes can be non-deterministic.

Based on the above, the dynamics of the bank account can be described both by which mode it is in (the symbolic dynamics) and how the money in the account changes over time (the 'quantitative' dynamics). There may be times where one is interested in the symbolic trajectories and times when one is interested in the 'quantitative' trajectories.

### 2.2.3 Compositionality and Feedback

A more complex dynamical system can be made by connecting two or more simpler dynamical systems. This is the idea of compositionality. A related idea is that of feedback. In the case of the composition of two systems, a feedback connection is such that at least one output of each system is an input to the other. This forms a loop as shown in top diagram in Figure 2.5. The 'loop' may not always be evident as shown in the bottom diagram in the figure. When more than two systems are composed, there are a number of possibilities: there may be a direct feedback connection between two blocks as shown in the top diagram in figure 2.6 or an indirect feedback path as shown in the bottom diagram.

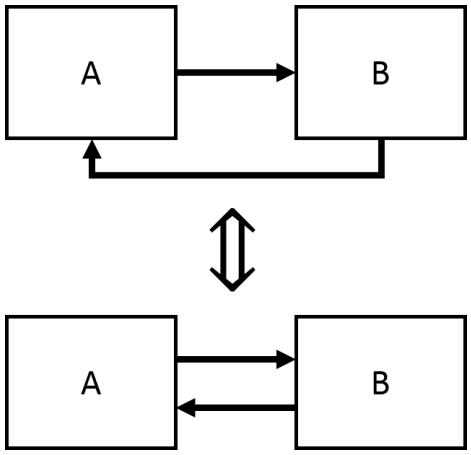



Figure 2.5: Feedback between two systems A and B. The two diagrams are equivalent

Feedback has many useful purposes, including influencing the dynamics of a system to make it more desirable and more robust to unwanted disturbances. It can also have

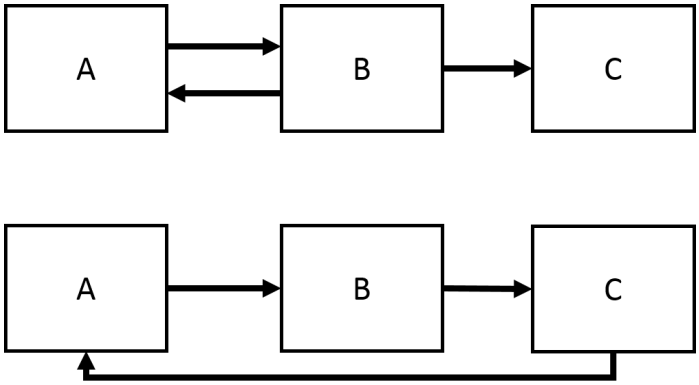



Figure 2.6: Feedback between three systems A, B, and C. The top diagram shows direct feedback between A and B. The bottom diagram shows an indirect feedback path from A back to A through B and C.

adverse consequences such as introducing undesirable dynamics. Most systems of interest are modeled as dynamical systems with feedback.

### 2.2.4 Intrinsic Behavior

The intrinsic behavior of a system describes how it responds to a variety of input trajectories in terms of state changes and producing outputs. If the system has some non-determinism, it also describes the nature of this non-determinism and whether it results in producing outputs that are not input dependent.

Continuing with the bank account examples, let's look at the behavior of an interest earning account that earns a monthly compound interest. The interest rate describes exactly how any amount that is in the account at the time the interest is applied is increased. The specific interest rate and the fact that the account is interest earning describe the intrinsic behavior of the account. In particular, a mathematical description of the account is

$$\begin{aligned} x[k+1] &= r \cdot x[k] \qquad k = 1 \ldots \\ \Rightarrow x[k] &= (1+r)^{k-1} x[1] \end{aligned} \tag{2.1}$$

where $x[k]$ is the money in the account at the end of $k$ months since the account was opened, $x[k+1]$ is the money in the account at the beginning of $k+1$ months since the account was opened and right after the interest has been applied to the previous month's amount, and $r$

is the interest rate (assumed here to be given in decimal form as a number between 0 and 1).

For any starting amount, if we know the interest rate, then the intrinsic behavior tells us exactly what would happen to that amount for a given input behavior. If we have two different accounts with two different interest rates, then starting with the same amounts and without any deposits or withdrawals, the one with the higher interest rate will have a higher amount in the account in the next month. Hence for the same input trajectories, two (deterministic) systems with different intrinsic behaviors will respond differently, and two systems with the same intrinsic behaviors will respond in the same way.

Some subtleties can arise. Since we are allowed to deposit and withdraw money, with the appropriate choice of inputs, we can make two accounts with different intrinsic behaviors exhibit the same state trajectories. Let's say one account has 10% interest and the other has 5%. If we do nothing to the one with 10% but at the beginning of every month, right after the interest has been applied, we deposit an amount equal to 5% of the amount before the interest was applied, then both should keep having the same amount of money in the accounts. Just looking at the states, it would appear that the accounts are the same, however if one considers the inputs to both accounts, we see that they are different. Hence two systems that behave the same way even when the input trajectories to each system are different must have different intrinsic behaviors.

To see why the behavior is called 'intrinsic', consider the number of months it takes to double the amount in the account. It turns out (with some simple manipulation) that this time (assuming we start from $k = 1$) is given by

$$k = \frac{ln(2)}{ln(1+r)} - 1 \tag{2.2}$$

Notice that this does not depend on the starting amount, but only on the interest rate $r$. Hence for *any* starting amount, it would take the same amount of time for the same account to double that amount. Remember that the intrinsic behavior is supposed to describe how the system responds to a variety of inputs. We can change the intrinsic behavior of the account either by changing $r$ keeping the base or general behavior (of compounding interest) the same, or we come up with a different way to handle the money in the account, resulting in very different general behavior.

## 2.3 Systems Safety Concepts

Systems safety concepts are best understood by looking at the operational view of the system as shown in figure 2.7. In general, a system is designed to achieve particular functional goals. This system is embedded in an operational environment. As the system operates and interacts with the environment, it produces a sequence of events ($[e_1,(e_2,e_3,e_4),\ldots,e_k,\ldots]$), some of which can be simultaneous (*e.g.,* $e_2$, $e_3$, and $e_4$). These event sequences are trajectories of the system, in this case more symbolic.

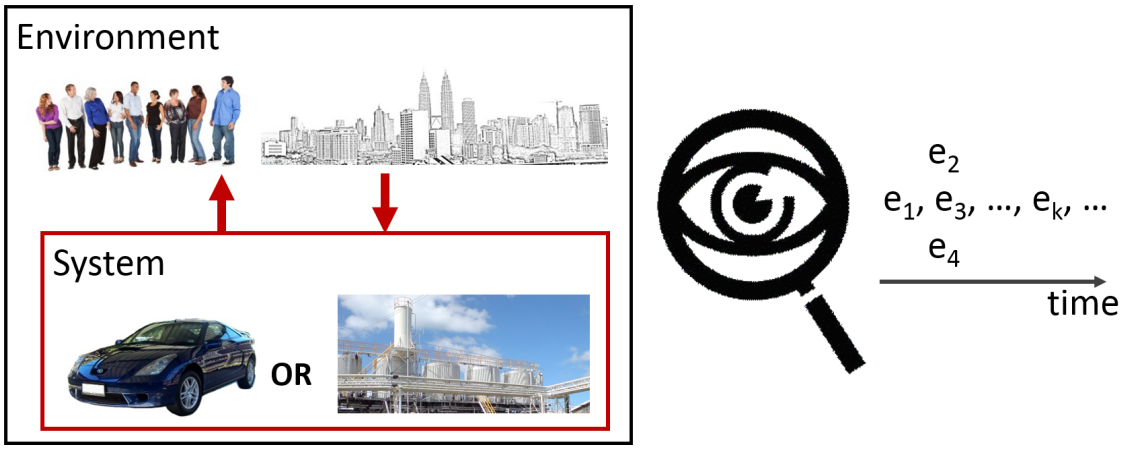



Figure 2.7: Illustration of operational view of the system from the perspective of system safety.

At the heart of systems safety is what are typically called mishaps or accidents. Generally, systems safety engineers agree that accidents are unplanned and undesired (series of) events that lead to some kind of loss. Loss could be loss of life, human function (due to injury) property, or finances. From the perspective of the illustration in figure 2.7, this means that we assume that we can decide whether any event results in a mishap or not (*i.e.,* we assume that in general there is some function or decision process $f_{\text{mishap}}$ such that for any event, $e_k$, $f_{\text{loss}}(e_k) \in \{\text{mishap}, \text{no mishap}\}$).

Another concept at the heart of systems safety is risk. Usually the concern is with accident or mishap risk. Risk is a combination of the likelihood of the event occurring and the severity associated with the mishap when it does occur. From the perspective of the illustration, if an event ($e_k$) results in a mishap (*i.e.,* $f_{\text{mishap}}(e_k) = \text{mishap}$), $p(e_k)$ is the

likelihood of that event occurring, and $q(e_k)$ is the severity of the mishap, then the risk is a function of these two quantities (*i.e.,* risk $= f_{\text{risk}}(p(e_k), q(e_k))$).

The aim of systems safety is to bring the risks to an acceptable level. A key activity in systems safety for doing this is the identification and mitigation of hazards. In general hazards are considered to be system states, conditions, or properties that under certain environmental conditions will lead to a mishap. Note that hazards or considered to be initiated within the system. Hence, they can be both introduced and controlled by design. Mitigating hazards involves both reducing the likelihood of occurrence and reducing the severity of the mishaps should they occur.

Different systems safety techniques differ mostly in the view of how mishaps occur (the paradigm or accident model). This in turn influences what is considered a hazard and how hazards are identified and controlled. They may also differ in which phase of the design process they concentrate on and whether they have been specialized for a particular industry or not. Below is a brief review of a number of systems safety paradigms and techniques. It focuses on the general characteristics of the paradigms and techniques with some examples; a comprehensive review is beyond the scope of this dissertation.

## 2.4 Systems Safety Paradigms and Techniques

The Safety Institute of Australia's (SIA) publication on "Models of Causation: Safety" [106] provides an informative account of different models of accident causation. The discussion below is based on their categorization of accident models, in which a model can be linear or non-linear. To aid in the discussion, the operational view illustration must be modified slightly to show the interaction of components within the system as shown in figure 2.8.

As mentioned previously, the different accident models drive the different techniques for identifying and controlling hazards. This an important point, since the choice of accident model will affect the way in which the systems safety engineering process will be carried out (*i.e.,* the way we would reason about accidents and safety). Hermitte, in a report reviewing accident models using in road accident research provides a useful discussion on this point looking at the history of accident models [43]. According to Hermitte, since in from around 1900 to 1920, accidents were thought to be random events, there was no need to look for causes. From the perspective of the illustration, this would mean that for

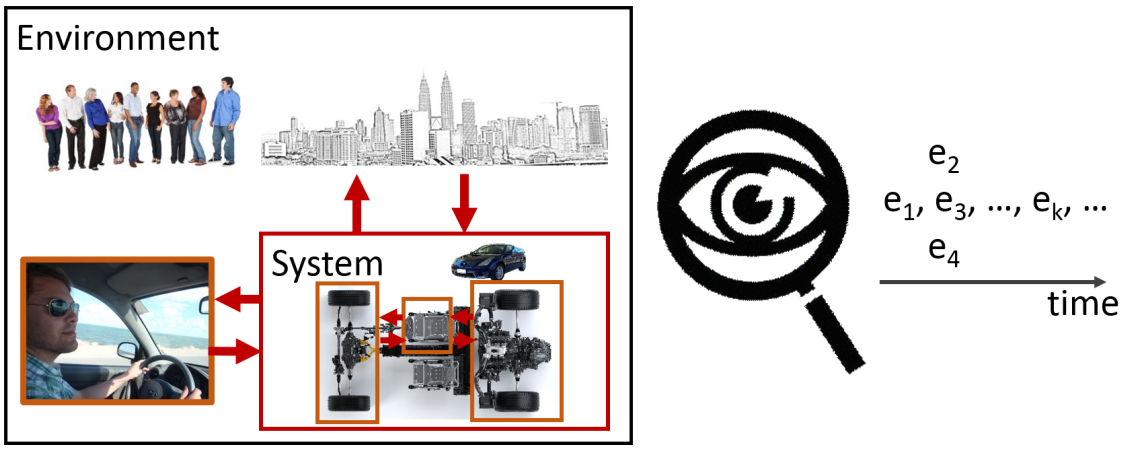



Figure 2.8: Illustration of operational view of the system from the perspective of system safety highlighting internal interactions.

any event ($e_k$) there is some probability that $f_{\text{mishap}}(e_k) = \text{mishap}$, which means that the trajectories of the system were random and non-deterministic.

From a little before 1920 to around 1940, it was believed that some drivers (those with personality disorders) were involved in others. He cites the reason (due to Hollnagel [45]) being that we classify people whose actions we cannot find plausible reasons for as psychologically disturbed, and since plausible reasons were still not available for accidents, then the drivers that caused them must be disturbed. From the perspective of the illustration, this would mean that if some event ($e_k$) results in an accident (*i.e.*, $f_{\text{mishap}}(e_k) = \text{mishap}$), then the driver shown must have a personality disorder. In dynamical systems terms, this meant that it was believed that drivers with personality disorders had intrinsic behaviors such that when put in driving situations, they would behave in ways the produced accidents.

It was only in the 1940s and 1950s that the idea that there were indeed (direct) causes for accidents and it was only finding them that would help prevent accidents. This idea of direct causation is what is embodied in the linear models category by the SIA. More recent models avoid the idea of direct cause altogether and take a more systems and behavioral view since accidents are complex and many have some non-deterministic aspects to them. Hermitte uses the example that even though driving under the influence increases the likelihood of an accident not every drunk driver gets involved in an accident. In addition, sober drivers do get involved in accidents. Hence, there are number of interacting factors (feedback loops) that result in an accident, which must be viewed in a more holistic manner. From

the perspective of the illustration, this means we must consider the behavior of the driver, the car, and the environment, and how the emergent behavior that results when they all interact could result in an accident. These more systems-based models are what the SIA terms non-linear models.

### 2.4.1 Paradigms and Techniques Based on Linear Models

In general, linear models follow what is called a "chain of events" paradigm. Form the perspective of the illustration, the idea is that if some event ($e_k$) results in a mishap (*i.e.,* $f_{\text{mishap}}(e_k) =$ mishap), then there are a series of events ($[e_i, \ldots, e_j]$) that preceded the accident that directly lead to the accident. Linear models are further categorized into simple ("single (root) cause") or complex ("multiple (latent) causes"). These models generally assume that the dynamical system has deterministic dynamics and hence accidents can be traced to initial conditions that eventually evolve into the accident.

#### The "Domino" Effect Model

The first and well-known simple linear model is the "Domino-effect" model by Heinrich [42]. Heinrich suggested that in general, the events leading to the accident (which results in an injury) is preceded by an unsafe act or a mechanical or physical hazard, which is preceded the fault of the person, which is preceded by the social environment or ancestry of the person. His contention was that the best way to prevent accidents was to focus on unsafe acts and mechanical and physical hazards. From the dynamical systems perspective, this model assumes that people have 'faulty' intrinsic behaviors that must be identified and reshaped to work well in the safety-critical context.

#### The "Swiss Cheese" Model

Another well-known (complex) linear model is Reason's Swiss Cheese model [88], in which an accident results when inadequate barriers are put in place to prevent it from happening. In this model, an accident is essentially waiting to happen, and the barriers are layers that prevent accidents from happen. When holes in these barriers develop and align, accidents "flow" through and happen. The holes in these layers are reminiscent of swiss cheese, hence, the name for the model. From the perspective of the illustration, internal to the system, events that result in accidents are being generated, but the system must be designed to filter these events out and only allow events that do not result in a loss to occur.

Reason's perspective is that human errors, which had been the focus of Heinrich's model, are inevitable, but must be managed and if not managed properly is what leads to accidents. The inadequate management is what has been called latent failure, and when combined with active errors (like human error), create holes in the defenses against accidents that allow events to flow through the holes that result in accidents.

This perspective in many ways has influenced the field of "human factors" which studies human-machine interaction and how to provide operating environments and user interfaces the minimize human error and are robust to some errors that would result in accidents. Aircraft displays are an example of where this influence has played a significant role.

Reason's approach, from a dynamical systems perspective takes the intrinsic behaviors of people (faults and all) as given, and rather tries to build a feedback systems around the people in the system to ensure that the inputs they receive and the outputs they produce help avoid overall system trajectories that would result in an accident.

#### Safety as Failure, Error, and Reliability Management

Model's like Heinrich's and Reason's have result in a number of approaches that we could term "safety as failure, error, and reliability management". A technique like Fault-Mode and Effect Analysis [97], for example, looks at the implications of faults and failure of components in the system as way of identifying potentially hazardous components (focusing on physical or mechanical hazards as dictated by Heinrich). Others like Fault-Tree Analysis [113] and Markov Modeling [85] focus on top-down chain of events, where multiple failures are thought to result in an accident. Markov modeling takes into account temporal ordering of failures.

In either case, the idea is that components inevitably fail or generate errors (a more Reason-like perspective) and if their reliability (probability of not failing at a point in time since beginning operation) is increased and other failure management techniques like redundancy are introduced, then the higher reliability will result in a safer system.

### 2.4.2 Paradigms and Techniques Based on Non-Linear Models

Non-linear models view the system as a whole, with components interacting to produce "complex (emergent) outcomes". The idea is that as system components interact, the events $[e_1, (e_2, e_3, e_4), \ldots, e_k, \ldots])$ emerge and if an accident results, it is due to an inappropriate interaction between components. The focus is not on failure (though failures are considered

as potential behaviors), since the interaction could have been due to a component behaving as specified. Whereas linear models, which tend to focus on failure, may believe that the system specification is correct with respect to safety, non-linear models assume that the specification itself could be wrong, and by considering emergent properties of the system, these potential flaws in specification can be identified and fixed. Non-linear models take a full dynamical systems view of the system.

### Functional Abstraction and Control of Risk

Rasmussen is credited with influencing the shift to the more systems-based approach use in non-linear models. In a paper in 1997 [87], he questioned the ability of prevailing models to deal with safety in what he called "a dynamic society." His focus was on the socio-technical aspects of safety and on the decision-making interactions between policy makers, systems designers, systems managers, personnel working with safety critical systems, and regulators that shaped the way the system was designed and eventually operated.

In particular, he proposed a shift from the structural decomposition approach typically used by linear models to a functional decomposition approach, where boundaries between functions are examined in order to use the relationship between constraints on the interactions at the boundaries and system performance to control safety. In essence, he was proposing a way to control the emergent behavior by controlling interactions: if emergent behavior was what resulted in accidents, then the hazards were related to the interactions, and by controlling interactions one could control hazards and reduce risk.

### Systems Theoretic Accident Model and Processes (STAMP)

In 2004, Leveson proposed a model of accidents reminiscent of Rasmussen's approach called the systems theoretic accident model and processes (STAMP) [59], and in 2011 released a book [60] based on experience with STAMP and tools developed based on it. Leveson actually credits Rasmussen (and others) for the perspective she adopts in developing STAMP. STAMP also focuses on interaction between components, but whereas Rasmussen was focused more on the socio part of the socio-technical relationship, Leveson's work elaborated more on the technical part.

In particular STAMP borrows ideas from control theory, where components have explicit roles as either a plant, controller, sensor, or actuator. The plant is typically the system to be designed which can exhibit behaviors that result (directly) in mishaps (*e.g.,* an airplane that can crash). The controller is a component that monitors the state of the plant

(through the sensor which provides it with information), decides on how to alter the state of the plant, and, if necessary, issues commands to the actuator to alter state of the plant.

Since the plant is what is typically thought of where accidents directly occur, STAMP and its associated tools focus on the idea of improper control. In particular its hazard analysis tool, Systems Theoretic Process Analysis (STPA) [58, 61, 77, 8, 105], focuses on identifying hazardous control actions. It is essentially a 'what-if' analysis focusing on what behaviors emerging if certain control actions are omitted or ignored, given too late or too early, or misapplied. It then focuses on causal factors that result in these hazardous control actions.

STAMP, however, requires the designer to identify what the accidents and hazardous states are. It mainly identifies how the system might transition into those states. It assumes that components have a known (and small) set of discrete states and actions. However, since it focuses on abstract systems behavior it is useful in the requirements stage of the design, and a new tool has been developed based on STAMP for early concept analysis [37]. Because it is based on the functional decomposition approach proposed by Rasmussen, it could also be used in later stages of the design as well.

### Formal Methods for Software-Intensive Systems

Recently, there has been interest in formal methods for dealing with complexity of software-intensive systems, especially for medical devices [47, 69, 46]. In these approaches, which originate from the computer science (and not systems safety) community, the focus is on assuring the correctness of the design of usually the decision-making software in a control system.

These approaches, similar to approaches like STAMP, require a specification of the hazardous states (or expected system behavior), the system design, and the behavior of its environment in a formal language. Techniques like model-checking [28] or theorem proving [92] are used to automatically (or semi-automatically in the case of theorem proving) verify that the design meets the requirements (and avoids hazardous states), and if issues are identified provide insights into the conditions under which the design fails to meet the requirements.

## 2.5 Systems Safety Principles

As different as the above approaches sound, they share some common principles. First, is that we must define what the accidents (mishaps) of concern are. Second, based on the accidents, and some model of how they occur, we must define what constitutes a hazard. Third, based on the hazards and our accident model, we must identify what aspects of the system to examine and modify in order to control the likelihood and severity of hazards.

The identification and control of hazards is usually iterative and continues all the way through operation of the system, but the earlier issues are identified the better. However, since no system is perfect, and we cannot anticipate everything ahead of time, we must decide on how much risk we are willing to accept and design to that risk level. This is where the definition of systems safety engineering provided earlier plays a role. All three principles come together to ensure that acceptable mishaps risks are achieved. Figure 2.9 illustrates the above principles

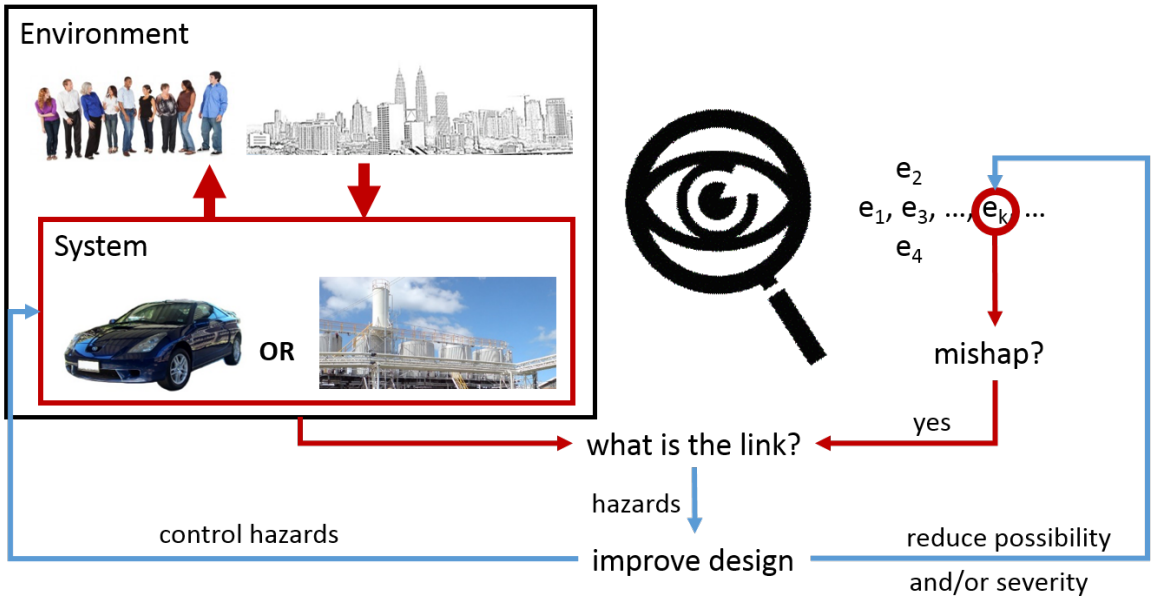



Figure 2.9: Illustration of the systems safety principles.

Safety-critical systems are all regulated, and typically a regulatory agency must decide if a design (or implementation) must be allowed to operate. Regulation may include continued surveillance of the system operation, but the crucial aspect is making the decision on

whether the system is allowed to operate or not. This creates an interesting scenario where the designer must argue that the design (and implementation) is safe based on inferences made from information gathered during the design process, and the regulator must decide whether to accept the argument made.

In many regulatory environments, there is guidance to designers on what factors into the criteria for an acceptable design. This is usually based on past experience. As complexity of systems grow and innovations inevitably appear, regulators must provide new guidance and develop new ways to reason about systems in order to make these pre-operation decisions. Such guidance finds its way into industry standards, sometimes focused on design features and sometimes focused on the design process.

In many industries, however, the hazards tend to be well-known and the need to identify new ones does not increase dramatically with innovations. The medical industry however, is not so lucky, with new devices addressing different health issues and hence different functions of the body, prompting a need to be able to develop criteria for new technologies as they arise and provide these criteria development mechanisms as part of guidance in addition to well-known hazards.

## 2.6 Applying Principles to Reasoning About Patient Safety

As mentioned previously, the current tools for systems safety grew out of systems where the functional goal of systems are not related to keeping humans healthy. However, in all safety-critical systems, human health is a safety issue (treated as a constraint to be balanced with functional goals). If human health is a safety issue, then for medical systems, where aiding in maintaining health is the functional goal, safety is the functional goal, and not a constraint. Hence the way we define mishaps must be explicitly tied to health. Many systems safety approaches have a limited treatment of health. Also, since previous medical devices were simpler, the safety burden was on the health practitioners, so safety considerations of these devices were quite limited.

Though newer efforts are under way to reexamine safety approaches for medical devices [57, 2], many are still grounded in the systems safety approaches where health is a constraint, and many focus on design and implementation issues. This is because they are being developed in the context of well-known medical systems where hazards have been identified over years. The hazards can therefore be taken as given. Confusion arises

with emerging technologies where situations are new and hazards have not been previously identified.

What is needed is a return to the systems safety principles to provide an approach for defining accidents, hazards, and acceptable mishap risk in a way that is valid for medical technologies and can be applied to emerging technologies. This is what this dissertation does. It defines mishaps in terms of health (which is cast as a safety issue) and links all other issues to this definition, providing a way for defining acceptable mishap risk which is both valid for medical technologies and consistent with systems safety principles. It adopts the non-linear approach, understanding accidents in terms of interactions and emergent system behaviors.

## Summary

Despite the diversity of systems safety techniques, they can all be considered variants of views of the evolution of a dynamical system, where the results of this evolution are intended to be safe (as defined for the particular scenario). Emerging medical technologies are stretching the applicability of these methods particularly because the pieces of what constitutes a hazard are not being revisited. These pieces are heavily influenced by the view of systems safety where health is constraint, and hence must be reexamined in the context where health is a goal.

The next chapter shows that using the dynamical systems view of human function, we can link many of the principles developed for systems safety to human function, and hence discuss health (and later patient safety) viewing the human a natural safety critical system. This allows us to take the necessary view of health management (where medical technologies are used) as aiding the body in maintaining safe states and to discuss patient safety in those health terms. In addition, it keeps patient outcomes at the center patient safety discussions while also allowing for discussion of design issues.

# Chapter 3

# A Model of Health Management

> You think because you understand "one" you must understand "two" because "one and one" makes "two". But you forget you must also understand "and".
>
> Jalal ad-Din Muhammad Rumi

## Chapter Overview

In the previous chapters, systems safety was introduced as a process of working towards achieving "acceptable mishap risk." In addition, the contention was that for medical technologies, the mishaps are undesirable patient outcomes. In order to agree on whether patient outcomes (mishaps) are undesirable (unacceptable) or not, we need some method to express what these patient outcomes are and how to interpret them.

This chapter develops a mechanism (a formal model) for doing this using dynamical systems ideas, creating a link between systems safety language and health management ideas. By showing the commonality of the dynamical systems view to both systems safety and health management stakeholders (practitioners, technology developers, and the FDA), it allows the patient safety model developed from it (in the next chapter) to be both valid from a health management perspective and consistent with systems safety principles. The model developed here therefore sets the stage for the rest of the developments in this dissertation.

## 3.1 Introduction

Health is about the elusive concept of acceptable human function. When a person has the flu, for example, the weakness caused by the virus prevents them* from doing the things they would normally be able to do otherwise. We generally consider this unacceptable function. Sometimes the concern is about such issues in the future. Being overweight may not prevent someone from doing what some would consider normal activities, but it can lead to cardiovascular problems which could eventually lead to premature death [71]. We can therefore say that health management is about addressing current or potential future loss of acceptable function.

There are a few implications of the above statements. First, we have to have some idea of what acceptable function is. Second, we must be able to identify the factors associated with current or potential future loss of function. Finally, we need a way of influencing the factors affecting a person's function to bring it to an acceptable level. The latter two parts of health management are not always possible, and the field of medicine has been dedicated to these since antiquity.

Ambiguity arises in all three statements: there may be disagreements on what is considered acceptable function, which factors affect function, and what is an appropriate course of intervention. One of the main aims of this chapter is to show that despite these disagreements, we can provide mechanisms (a language, if you will) for expressing ones conception of these ideas precisely. This enables different conceptions of any or all of these ideas to be discussed in a coherent manner, and hopefully leads towards consensus or at least an understanding of the differences. Being careful about the ambiguity of health is important, since one of the aims of health management is to intervene when we view function to be unacceptable. Intervening when one should not can prove to be harmful.

The example of diabetes and diabetes management below sums up the above discussion concretely. The example is general, but will be expanded as needed later to help illustrate other concepts.

**Example 3.1:** Diabetes and Diabetes Management

Diabetes is health condition where the body's ability to keep blood glucose levels down is impaired [73]. These can lead to complications such as blindness, loss of

*"They" and its derivative forms is used as the gender-neutral third person singular pronoun throughout this dissertation.

sensation in the legs, and other complications that lead to poor quality of life and eventually death [22]. Very high blood glucose can also have severe effects [22]. These complications are the unacceptable loss of function.

The blood glucose level is the factor affecting this unacceptable loss of function. Based on this connection between blood glucose and these complications, one could say that the body's inability to keep blood glucose levels down is also unacceptable function.

Proper diet and exercise can be used as an intervention in the case where the body retains some ability to react to high glucose levels (in the case of Type 2 diabetes). When the body's ability to react to high glucose levels is completely impaired (in Type 1 diabetes), infusing or ingesting insulin (a hormone that helps reduce blood glucose levels) is used as the intervention and part of the management strategy.

One source of ambiguity is the list of things that are considered unacceptable loss of function. Blind people, with the right accommodations, do function quite 'normally'. Certainly, there are activities that they may not be able to partake in feasibly, but whether their condition is "unacceptable" or not can become a touchy issue. Some, however, would agree that losing one's sight because of a condition like diabetes is unacceptable.

In addition, though standard limits exist for when a person is considered diabetic or not [10], the actual level at which it is a problem for a particular person can vary from person to person. The World Health Organiztion report on defining and diagnosing diabetes [118] provides na informative discussion on the issues with defining such limits.

This dissertation uses the language of dynamical systems to cast health and health management in more precise terms and reduce the ambiguity in expressing our conception of health. The language of dynamical systems makes treating health as a safety issue and hence safety as the functional goal of health management systems natural, as is shown below.

## 3.2 Humans as Dynamical Systems

Humans can be viewed as dynamical systems. We take in inputs in the form of various stimuli from the environment (*e.g.,* food, light, temperature, social contact with other humans). We produce outputs in the form actions on our environment (*e.g.,* moving around,

exhaling air) or in the form of quantities observable by cognitive processes or by the use of instruments (*e.g.,* organ structure, speech, 'body language'). We react to inputs to change state (*e.g.,* eating carbohydrates increases blood sugar) and produce outputs (*e.g.,* hot environments induce sweating, an insult may result in a fight). We have parameters (*e.g.,* age, height, weight, size of heart, strength of muscles, insulin sensitivity, genetic makeup, personality traits) that tend to govern our evolution (*e.g.,* taller people move further in fewer steps than shorter people, introverts tend to avoid certain social situations). Human parameters also evolve (*e.g.,* we age, we grow, we gain and lose weight).

The are many aspects to health management (*e.g.,* physiologic, biomechanical, mental, behavioral, cognitive). Thankfully, a dynamical systems view of humans is not foreign to many of these areas. The most commonly encountered models are of physiologic processes [63, 65] and of human movement [115]. However, there are also models of behavior [89] and of cognition [91]. In general, a person is in constant interaction with their environment so our model of humans as dynamical systems[†] captures this as shown in figure 3.1.

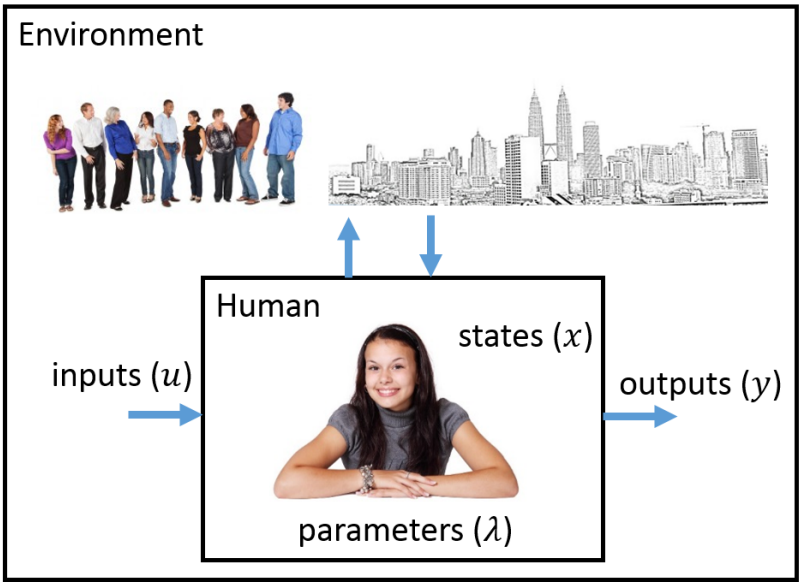



Figure 3.1: Human as a dynamical system.

[†]We use the term "patient" or "person" to denote the human dynamical system consisting of the person in their environment in the rest of the dissertation.

## 3.3 Health as a Measure of Human Function

In order to arrive at a notion of acceptable human function, we need some way of measuring human function. In our diabetes example, one function we are concerned about is how well blood glucose levels are controlled by the body. The blood glucose level can be thought of a state variable. In order to determine how well it is controlled, we would observe how it changes over time under different conditions. In essence, we are concerned with the trajectory of the blood glucose level. The trajectories of concern need not be limited to state variables; in some cases, it is more convenient to consider an output or a parameter. The method of analysis of the trajectories of interest is called the health metric and the result of the analysis is called the outcome. These notions are formalized below

**Notion 1:** Health Metrics and Outcomes

A health metric ($\mu_H$) is a mapping from a set of finite sequences of trajectory values to a finite set of 'objects' suitable for comparison. The concept is best described through the illustration in Figure 3.2.

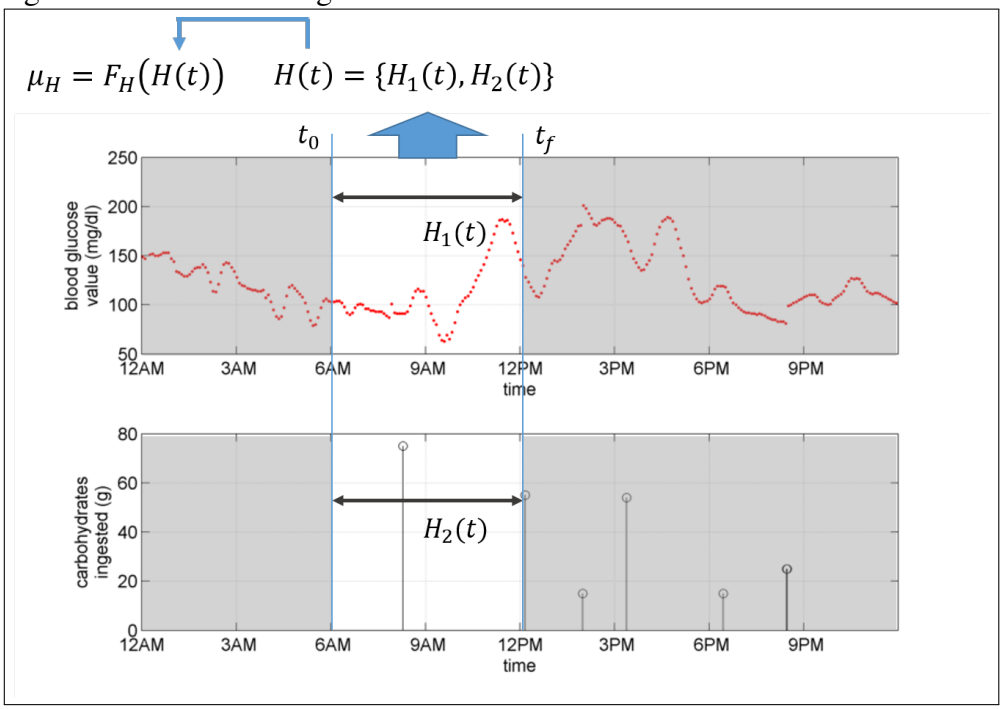



Figure 3.2: Illustration of the health metric notion

$H_1(t)$ and $H_2(t)$ (the values in the unshaded portion) are the finite sequences of trajectory values, and $H(t)$ is the set that contains them. They are finite sequences because the values (in the unshaded portion) lie between the defined finite time range $[t_0, t_f]$. Each trajectory value sequence $H_i(t)$ can be related to inputs, outputs, states, or parameters. In this particular case, $H_1(t)$ is related to a state trajectory (the blood glucose value), and $H_2(t)$ is related to an input trajectory (the meals that the person takes). In general, if a sequence is related to states or parameters, they would be considered estimates since these are usually not considered directly observable.

Notice that the trajectory can have no values for a significant part of the time window or can contain a single value like the case of $H_2(t)$. In general, the set of trajectories need not be uniform in any way. The only requirement is that the values considered fall within the particular time window of interest.

The health metric map ($F_H$) is computed on this set of trajectory sequences to get a value which would be the health metric outcome. Summarizing the above, this means that the health metric is formally given by

$$\begin{aligned} \mu_H &= F_H(H(t)) \\ H(t) &= \{H_1(t) \ldots H_n(t)\} \\ H_i(t) &= [h_i(t_k) \ldots h_i(t_m)],\ t_0 \leq t_k < t_m \leq t_f,\ m \geq 1 \end{aligned} \tag{3.1}$$

The outcome of the health metric ($\mu_H$) can be any n-dimensional object. It could be anything from a single value to a sequence, a probability distribution, or a collection of these and other objects. The choice depends whoever defines the health metric. Simpler objects are easier to manipulate and use in later analysis.

Lastly, the description of the health metric must also capture the context. The context can be described in terms of time (the way $[t_0, t_f]$ is defined), or in terms of the nature of other trajectories, or both. In the example above, a context based on time would mean that the health metric requires blood glucose and meal information between 6am and 12pm. If the context is based on a trajectory, say the meal, then it could be the metric depends on blood glucose values in a six-hour period when only one meal is taken. If it is based on both, then it could be that the metric depends on blood glucose values between 6am and 12pm when only one meal is taken.

Before providing some examples of how the above notion applies, there is one more issue to address and notion to introduce. The above developments tell us what defining and

computing a health metric entails, but we are still with the question of what the outcome of a health metric indicates. One thing we do know is it must be related to some notion of acceptable or unacceptable with respect to human function.

Let's assume that the dynamics of a person can be categorized as having two general modes, acceptable ($a$) and unacceptable ($\neg a$) as shown in figure 3.3. Remember that the modes of a dynamical system have distinct dynamics which can be characterized by the trajectories. Since the health metric is a measure (and hence characterization) of trajectories, it can, in effect, indicate the mode of the person's dynamics. The first thing the health metric therefore tells us is whether the person is currently functioning in an acceptable or unacceptable mode ($\mu_H = F_H(H(t)) \Rightarrow a \oplus \neg a$).

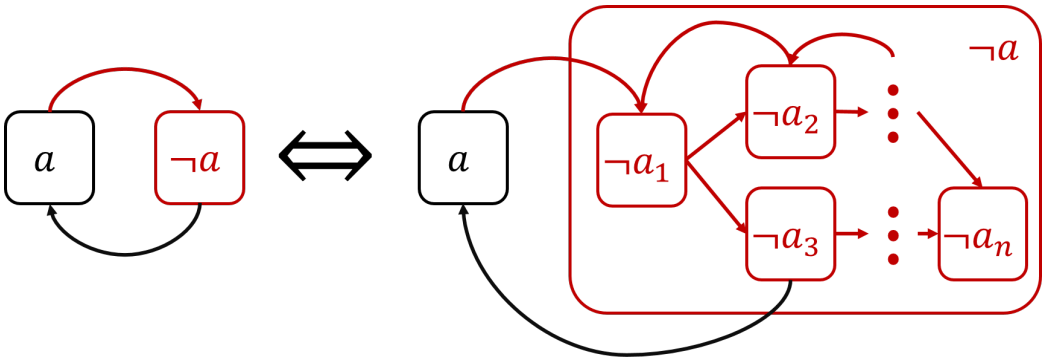



Figure 3.3: Illustration of two modes of human dynamics. $a$ is the acceptable mode and $\neg a$ is the overall unacceptable mode. Modes of the form $\neg a_{i\in\{1,2,3,n\}}$ are the unacceptable submodes

The second thing the health metric outcome tells us is how unacceptable the current dynamics of the person is. Let us further assume that within the unacceptable mode are submodes of unacceptability ($\{\neg a_1 \ldots \neg a_n\}$) as shown in figure 3.3 with some being worse than others (*e.g.,* being dead might be considered worse than being alive and with a cold). In this case, the health metric tells us about the likelihood of transition between this unacceptable submodes ($p(\neg a_i \rightarrow \neg a_j)$).

In addition, this likelihood is usually qualified by a time horizon within which transitions might occur, as we will see in the examples below. Hence, some metrics indicate a likelihood of transitioning to the other submode in short period of time (on the order of seconds, minutes, hours, or days), whereas others indicate a longer period of time (on the order of weeks, months, or even years). There may be a likelihood of transitioning from one of these unacceptable submodes to the acceptable mode, but because these submodes retain a

likelihood of transition to other unacceptable submodes, they are considered unacceptable. The above is formalized in the notion below.

**Notion 2:** Health Metric Outcome Interpretation

The health metric outcome can be interpreted on a symbolic scale of acceptable or unacceptable. Assuming that a outcome is scalar value, for simplicity, the scale can be visualized in the form of the graph in figure 3.4. On the x-axis is the value of the outcome ($\mu_H$) and on the y-axis it the symbolic scale of acceptability. The graph shows the mapping between the outcome value and acceptability.

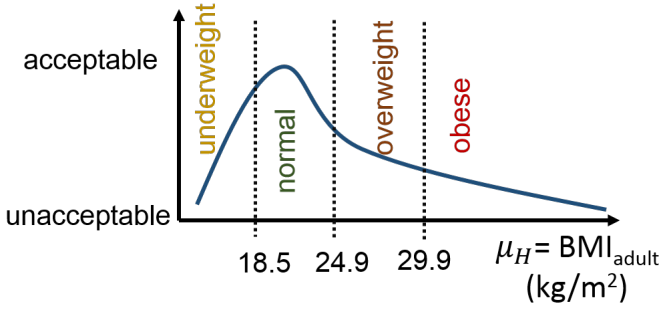



Figure 3.4: Example of mapping of health metric outcomes to an acceptability scale using body-mass index (BMI). Data obtained from Centers for Disease Control and Prevention website: http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

In this case, the health metric is the adult body-mass-index (BMI). In general, the BMI is based on the height, weight and age of the person. For adults, no adjustments are made for age or gender, but in order to use the adult version of the metric and scale, the person must be 20 years or older [24]. Height and age can generally be considered as parameters. It is debatable whether weight is a parameter or state. Nevertheless, the formula for the adult BMI in our framework (assuming weight is a parameter) is

$$\begin{aligned} \mu_H &= F_{BMI_{adult}}(\{\lambda_{height}(t), \lambda_{weight}(t), \lambda_{age}(t)\}) \\ &= \frac{\lambda_{weight}(t_k)}{(\lambda_{height}(t_k))^2}, \ \lambda_{age}(t_k) \geq 20\text{years} \end{aligned} \tag{3.2}$$

The height and weight must be in SI units (*i.e.,* meters (m) and kilograms (kg) respectively). Notice that for this metric, the age parameter only provides context (*i.e.,* it is

not an adult BMI and we cannot use this interpretation scale unless the person is more than 20 years old).

This method of interpretation is not restricted to scalar-valued health metrics. Since it is a mapping, one can be created for arbitrary metrics, though the process could be quite complicated. Scalar or n-dimensional quantitative metrics are certainly much easier to deal with than more complex mathematical objects.

One thing to note is that the time scale for transition to further unacceptable states is not shown on the graph. The graph is only a visualization so the scale need not be defined as a graph. One must, however, state in the definition of the acceptability scale what the time scale of the likelihood of transition to further unacceptable states is in order to reduce any ambiguity with what the health metric indicates. For BMI, the concern is with long term risks (on the order of years)[24].

One ambiguity that needs to be resolved is what the acceptable state is. One answer that seems satisfactory is to say that an acceptable state is one where if those dynamics are continued (or repeated) does not create the likelihood of transitioning to an unacceptable mode, or keeps such a likelihood very low. This implies that the dynamics of an unacceptable submode intrinsically creates or increases likelihoods of transitioning into further unacceptable submodes. This also implies that transitions from the acceptable mode to unacceptable submodes are triggered by non-deterministic actions.

The examples below put the above ideas in more concrete terms.

**Example 3.2:** Short-Term Considerations for Blood Glucose Values

A single blood glucose measurement at any point in time can be used to determine short-term consequences, which usually can be fatal. Both very high levels [75, 48] and very low levels [30] can result in death. In either case, the concern is based on a single measurement (or the average of a very small sample of repeated measurements within a short time-frame). Using our health metric formalism, a health metric based on blood glucose for short-term considerations would be

$$\begin{aligned} \mu_H &= F_{BG_{st}}(\{x_{BG}(t) = [x_{BG}(t_1)\ldots x_{BG}(t_N)]\}) \\ &= \sum_{k=1}^{N} \frac{x_{BG}(t_k)}{N} \end{aligned} \tag{3.3}$$

where $x_{BG}(t_k)$ is a sample of blood glucose level at the time $t_k$, $N$ is small (less than 5), and $t_n - t_1$ is on the order of a few minutes.

In this case, we are interested in extreme values (either too high or too low), and we can interpret the health metric on scale of acceptability as shown in Figure 3.5. As before, the x-axis shows the outcome of the metric and the y-axis shows the scale of acceptability (lower on the axis indicates more unacceptable). The graph shows the mapping of outcomes to a level of acceptability.

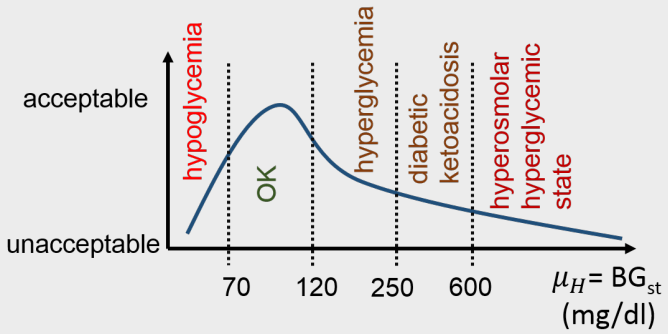



Figure 3.5: Mapping of blood glucose value to acceptability scale for the purpose of short-term considerations.

In the case of diabetic ketoacidosis and hyperosmolar hyperglycemic state, additional information is needed to confirm the condition, though blood glucose levels are a good indicator.

**Example 3.3:** Long-Term Considerations of Blood Glucose Dynamics

Fasting blood glucose is used as an indicator of how well the body controls blood glucose. The fasting blood glucose metric is characterized by a sample of blood glucose usually more than 8 hours after any caloric intake [10]. The visualization of the inputs to the metric are shown in Figure 3.6.

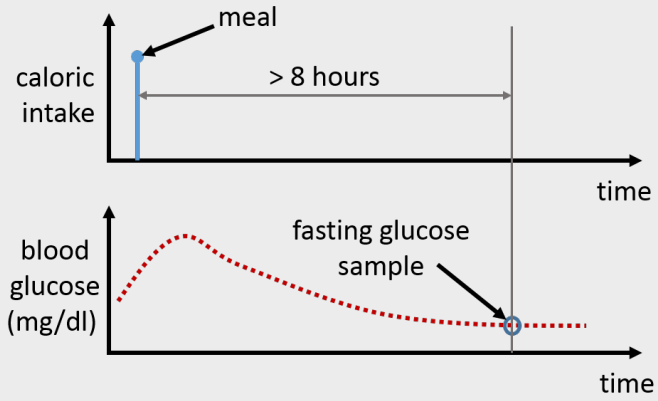



Figure 3.6: Visualization of fasting glucose health metric.

Here the meal trajectory only provides context. It essentially states that the meal trajectory must be similar to the one shown in order for the metric to be correct (*i.e.,* having a meal closer to when the sample is taken would not make the outcome value a fasting blood glucose sample). Note that although the whole blood glucose trajectory is shown, only the indicated sample is actually observed. What happens before and after the sample is taken is unknown to the person computing the metric.

The fasting glucose is a good indicator of blood glucose control because it is expected that after that amount of time after a meal, the body should be able to bring the blood glucose down to normal levels and if that is not the case then it indicates issues with the mechanisms for controlling blood glucose which would indicate that the person is diabetic [10].

Another similar metric is the oral glucose tolerance test which tests the body's response after two hours to a known amount of glucose intake (equivalent of 75g of anhydrous glucose dissolved in water) after at least an 8-hour fasting period [10]. To get a more complete picture, the World Health Organization (WHO) recommends using both metrics [118]. This results in a two-dimensional overall metric which is given by the visualization in Figure 3.7. The acceptability scale for this two-dimensional metric is given by Figure 3.8

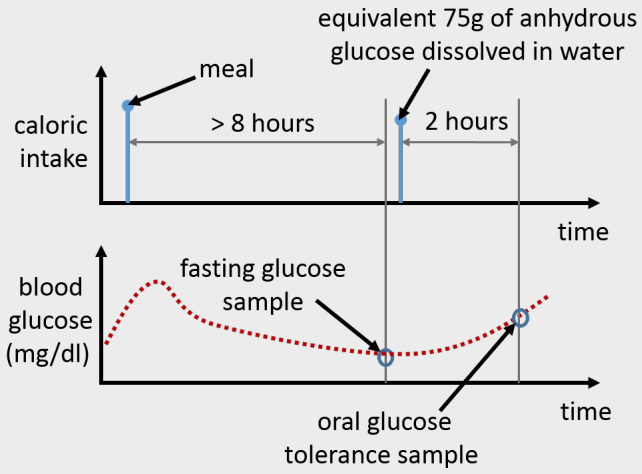



Figure 3.7: Visualization of combined fasting glucose and oral glucose tolerance test health metric.

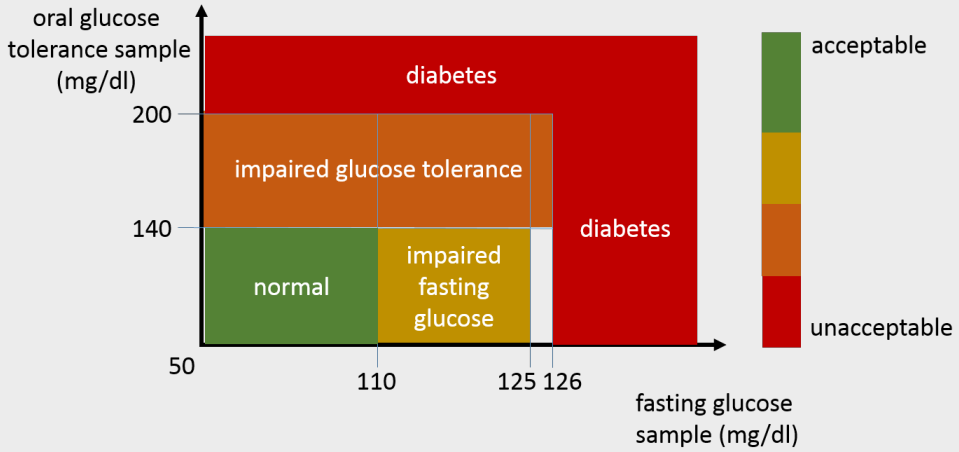



Figure 3.8: Visualization of combined fasting glucose and oral glucose tolerance test health metric mapping to acceptability scale. The scale is based on information provided in the WHO publication on "Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia" [118].

The following examples are intended to demonstrate the generality beyond physiologic variables of the health metric notion. It looks at the body-mass index, analysis of biopsied

tissues, and the patient health questionnaire-9, representing three different perspective of human function. Body-mass index is concerned with physiological issues, particularly risk of cardiovascular disease. Biopsies are used to detect potentially cancerous tissue which is an anatomical (structural) issue, that can affect physiologic function. The patient health questionnaire is focused more on behavioral and mental health.

**Example 3.4:** Body Mass Index

We saw body mass index as health metric in the the discussion of interpretation of health metric outcomes. There were looked at adult BMI, which is essentially a one-dimensional metric, since it is based only on the value of the output BMI formula. The interpretation for BMI for children (aged 2 years to 20), however depends on the BMI formula (given by equation 3.2), the gender, and the age of the child. Figure 3.9 shows the interpretation of BMI for boys with an example interpretation for a 10-year old. Notice how the weight categories change with age. From this perspective, the BMI for children has a two-dimensional output value (the BMI calculation and the age).

The age adjustments are actually based on the percentiles of boys with the particular BMI value. Underweight represents those below the 5th percentile, normal weight are those who fall between the 5th and 85th percentile, overweight are those within the 85th and 95th percentile, and obese are those above the 95th percentile.

**Example 3.5:** Biopsied Tissue Analysis for Cancer Testing

When a growth is suspected to be potentially cancerous, doctors perform a biopsy to confirm suspicions. The biopsied tissue is examined by a pathologist under a microscope. Based on how "normal" the cell looks, it is assigned a tumor grade. "Normal" cells are those that have specialized for the particular function in the particular area where they are found. Abnormal cells are less specialized. The microscope analysis shows how normal or abnormal a cell is. Less specialized cells are more likely to keep dividing and spreading, making them more cancerous, whereas more specialized cells are less likely to do so. The general grading scheme is on a scale on 1 to 4 (1 being less likely to grow and spread and 4 being most likely) [70].

From our health metric formalism, the biopsy analysis is the health metric. It's input is the state of a cell at the time of biopsy (which includes the structure of the cell). The output is the tumor grade.

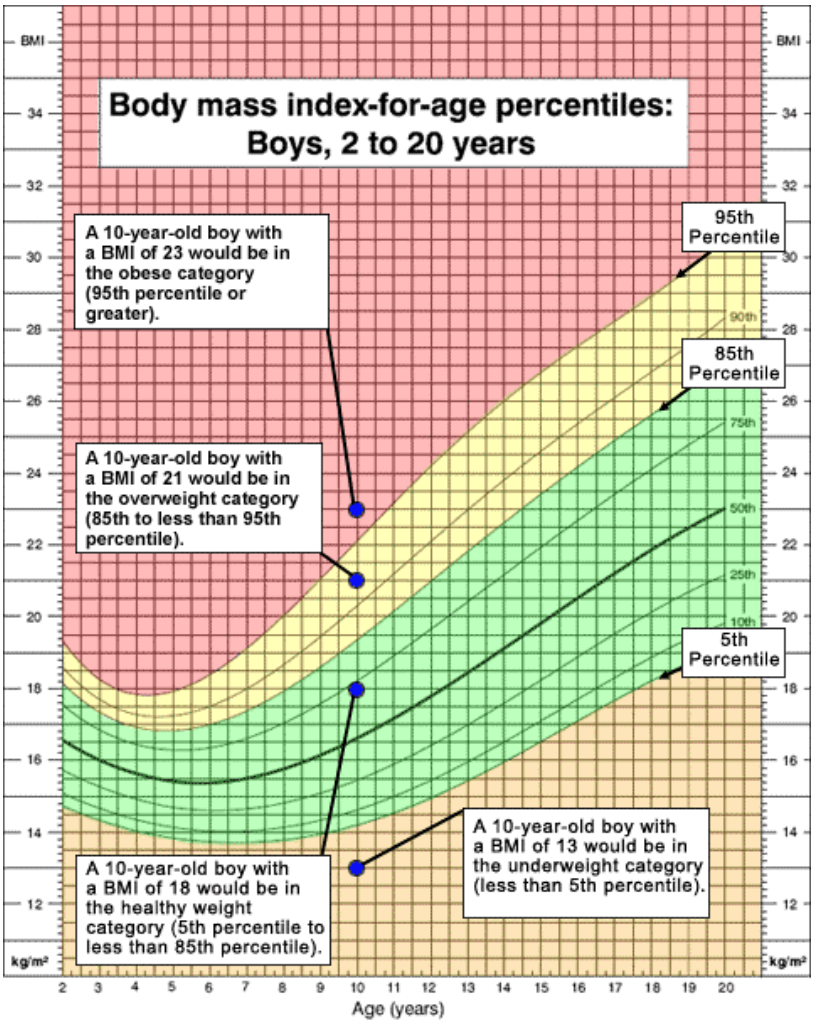



Figure 3.9: BMI interpretation for boys showing the interpretation for a 10-year old. Source: Centers for Disease Control (url: http://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html [Last accessed: April 6, 2015])

**Example 3.6:** The PHQ-9 Measure for Depression

The patient health questionnaire-9 (PHQ-9) is a questionnaire [96] administered to patients in order to assess the severity of depression [53]. The questionnaire is shown in figure 3.10. The questions are based on the patient's observations over the previous two weeks. The output is a score (on a range of 0 o 27).

| **Over the last 2 weeks, how often have you been bothered by any of the following problems?** *(Use "✔" to indicate your answer)* | Not at all | Several days | More than half the days | Nearly every day |
|---|---|---|---|---|
| **1.** Little interest or pleasure in doing things | 0 | 1 | 2 | 3 |
| **2.** Feeling down, depressed, or hopeless | 0 | 1 | 2 | 3 |
| **3.** Trouble falling or staying asleep, or sleeping too much | 0 | 1 | 2 | 3 |
| **4.** Feeling tired or having little energy | 0 | 1 | 2 | 3 |
| **5.** Poor appetite or overeating | 0 | 1 | 2 | 3 |
| **6.** Feeling bad about yourself — or that you are a failure or have let yourself or your family down | 0 | 1 | 2 | 3 |
| **7.** Trouble concentrating on things, such as reading the newspaper or watching television | 0 | 1 | 2 | 3 |
| **8.** Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual | 0 | 1 | 2 | 3 |
| **9.** Thoughts that you would be better off dead or of hurting yourself in some way | 0 | 1 | 2 | 3 |

Figure 3.10: The questionnaire for the PHQ-9. Source: http://www.phqscreeners.com/pdfs/02_PHQ-9/English.pdf

Figure 3.11 shows the interpretation of the score and the recommended course of action. From the perspective of our health metric formalism, the PHQ-9 produces a one-dimensional output value. Its input is based on behavioral and emotional trajectories within a two-week period. The metric is the questions, checking the frequency of occurrence of certain feelings or behavioral situations.

| PHQ-9 Score | Depression Severity | Proposed Treatment Actions |
|---|---|---|
| 0 – 4 | None-minimal | None |
| 5 – 9 | Mild | Watchful waiting; repeat PHQ-9 at follow-up |
| 10 – 14 | Moderate | Treatment plan, considering counseling, follow-up and/or pharmacotherapy |
| 15 – 19 | Moderately Severe | Active treatment with pharmacotherapy and/or psychotherapy |
| 20 – 27 | Severe | Immediate initiation of pharmacotherapy and, if severe impairment or poor response to therapy, expedited referral to a mental health specialist for psychotherapy and/or collaborative management |

Figure 3.11: The PHQ-9 interpretation scale. Source: http://www.phqscreeners.com/instructions/instructions.pdf [p.7], which reproduces it from the paper by Kroenke and Spitzer where the PHQ-9 was introduced [54].

From the above, some links can be drawn between the dynamical systems view of systems safety, and that of health management. For one, the health metrics act as indicators of risk. They indicate the probability of certain modes related to mishaps (whose severity we are aware of and want to prevent) occurring. In fact, the term "risk" is used quite often in discussing health metrics. We can therefore say that the human body is a natural safety-critical system and health is about maintain safe states. The health metrics then indicate the presence of hazards. When the risks are acceptable, we allow the body to function on its own, and when they deemed unacceptable, they must be controlled to reduce them. This control is the job of the health management system as discussed below.

## 3.4 Health Management as a Feedback System

When health metric outcomes are unacceptable, we intervene, if possible, to restore outcomes to acceptable levels. This is done by designing a system ($M$) consisting of a sensing subsystem ($S$) to observe and infer the current situation, a decision-making subsystem ($D$) decide on appropriate course of influence, and an actuation subsystem ($A$) to influence the person or their environment.

This system[‡] can contain both humans (including the person whose dynamics we intend to influence) and machines in any of these subsystems. It interacts with the patient, and their environment, in feedback fashion (as shown in Figure 5) with the intent that the evolution of the overall feedback system results in the desired evolution for the person, which produces health metric outcomes that indicate lower risk.

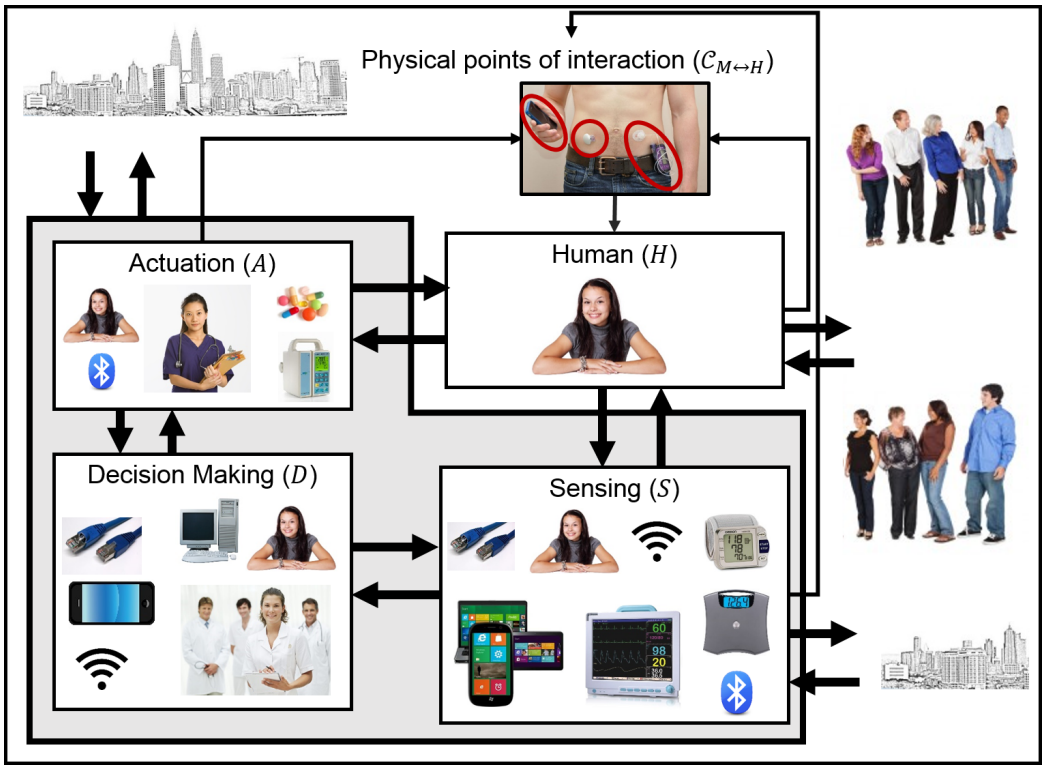



Figure 3.12: The health management system interaction with the person in feedback fashion.

Note that the parts that interact directly (physically) with the human system are the sensing and actuation subsystems: the decision-making subsystem interacts indirectly with the human through the sensing and actuation subsystems. The physical points of interaction ($\mathscr{C}_{M\leftrightarrow H}$) are important because the human is a non-uniform system exhibiting spatio-temporal dynamics. An example of this importance is lead placement issues for ECGs [41].

[‡]the term "system" to refers to the health management system.

Also note that the interactions with the human are feedback interactions. Typically, sensing it thought of as a one-way interaction from the human to the sensing subsystem, and actuation as a one-way interaction from the actuation subsystem to the human. However, the sensing subsystem can affect the human because of the direct interaction between them. For example, taking an X-ray requires irradiating a person even though the radiation is not meant to be therapeutic. Likewise, the human can affect the actuation subsystem. For example, a patient on a intravenous fluids can remove the needle used to infuse fluids.

In addition, the system interacts with the person's environment. These are sometimes useful (*e.g.,* a doctor gathering information from family members), but can also interfere with the work of the system (*e.g.,* a mobile system affected by excess heat or humidity in the person's environment).

Lastly, the internal subsystems of the health management system interact with each other in feedback fashion as well, sometimes in complex and unexpected ways. In addition, these subsystems are made up on components that also interact in feedback fashion. Furthermore, there are number of indirect interaction paths; for example, the sensing subsystem interacts indirectly with the person through the decision-making and actuation subsystems.

All these interactions will affect how the overall system evolves and hence how the person's dynamics evolve. It is this evolution of the whole system that we are concerned with when discussing patient safety. In terms of the system design in particular, we are interested in how the characteristics of the health management subsystems and the nature of their interactions (both with each other and the person whose dynamics we are trying to influence) affect the dynamics of the whole system, and hence the dynamics of the person.

## 3.5 Relation to other Health Management Modeling Work

As mentioned previously, modeling human function and health management from a dynamical systems perspective is not a new concept. Health fields take this view of humans and health [23, 63, 89, 91, 115, 99], and the interest in doing so continues to increase [89, 63, 99]. Interestingly, a number of significant advancements in the field of dynamical systems were made in the 1940s by Norbert Weiner and his colleagues who were developing mathematical theories for physiology [116], and a relatively recent article by Carpenter [23] claims that this recent interest is only a rediscovery of what physiologists knew decades ago.

In our case, however, this history of dynamical systems in health allows us to establish the validity of the model developed in chapter 3 based on our dynamical systems modeling of health. In addition, the focus of the model presented in this chapter is on how we identify what is acceptable and unacceptable function (*i.e.,* how we measure health). Much of the previous and current work in modeling human function using dynamical systems has been to understand mechanisms for normal function or those surrounding undesirable outcomes. Some models are also geared and developing and evaluating treatment strategies for cases where interventions are needed [29, **?**, 65, 89].

Our modeling, on the other hand, provides insights into the general structure and interpretation of health metrics, and can serve as the basis of developing health metrics for emerging technologies. Some development of metrics has been done for emerging technologies, for example, the low blood glucose index and average daily risk range established by Kovatchev *et al.* [50, 51]. The treatment here is focused more of a general way to derive and interpret metrics. This precise view of health metrics also allows more meaningful discussion of patient safety based on health metrics, since the built-in assumptions and the rationale for the choice of what is acceptable and unacceptable is made explicitly, opening these up for discussion. I am not aware of any such general modeling of health metrics.

## Summary

The way we view health is central to the way we manage it. How we measure human function and decide it is acceptable or not determines when and how we choose to intervene if possible. Because of the ambiguity of health, disagreements on these points can lead to confusion on what is acceptable and unacceptable function and how to intervene. This chapter introduced and general and precise mechanism for describing the way we reason about health called health metrics.

These metrics, based on the dynamical systems view of human function, simply measure function and indicate either the possibility of immediate or future detrimental consequences. Health metrics can be simple or complex, they require information from the person of interest obtained by human observation or the aid of medical technologies. They are also contextual, and this chapter showed that all these could be described using the idea of trajectories of the human dynamical system.

From the perspective of the discussion of systems safety in chapter 4 health metrics are the risk indicators (*i.e.,* they indicate the possibility of a loss (or an event leading to it)

occurring some time in the future). Health management therefore constitutes introducing a system to help the body keep these metric values to levels where risks for detrimental consequences are reduced. In systems safety terms, this is analogous to risk control by reducing hazards. The next chapter looks at reasoning about patient safety of medical technologies used in health management from this health metric perspective. This puts the patient outcomes at the center of our patient safety discussions making our developments valid from health management perspective, while keeping them consistent with systems safety principles.

# Chapter 4

# A Model of Patient Safety

> Methods and ways of looking are not propositional in character. They are not true or false, and we create great confusion for ourselves if we try to assess them in that way or mistake a method of analysis for a theory, a proposition or anything else two-valued and testable.
>
> Guy Robinson

## Chapter Overview

Chapter 3 developed the systems safety of "acceptable mishap risk" in the context of human function, independent of medical technologies (*i.e.,* the case where the person does not interact with the health management system). This chapter builds on those ideas to develop notions for defining "acceptable mishap risk" (the patient safety criteria) in the context of health management. The notions together constitute the formal model of patient safety.

Risk arises in health management because coupling the health management system to the human dynamical system can result in undesirable outcomes. This chapter provide notions for accounting for these undesirable outcomes (and the reasons for them) in reasoning about patient safety. The validity of the model is demonstrated through a case study. Two case studies are used to demonstrate its utility for aiding in various aspects of reasoning about patient safety of medical technologies.

## 4.1 Introduction

Coupling a health management system to the human dynamical system as discussed at the end of chapter 3 will result in three possibilities: (1) the interactions produce dynamics in the human that are acceptable; (2) the interactions fail to correct the unacceptable dynamics that prompted the introduction of the health management system; (3) the interactions produce new unacceptable dynamics which did not exist before the introduction of the health management system (what are known as "side-effects"). The second and third possibilities are the concern of this chapter.

There are four main reasons why the health management system might fail to achieve its functional goal or introduce side-effects. The first is that from a dynamical systems perspective, the human body is a multiple-input, multiple-output (MIMO) system. For such systems, one input can affect multiple states and outputs. Hence, even though a health management intervention may intend to affect only a certain subset of these variables, it may invariably affect other variables causing side-effects. A physiologic model like Hummod [44] provides this MIMO perspective of physiologic function.

The second and third reasons are both related to the variability that the human body exhibits which causes non-deterministic behavior. For the same person, the human body consistently changes both structurally and physiologically. In addition, as mentioned before, the human body has its own control mechanisms for maintaining its function (what is known as homeostasis in physiology). These controls adapt to damage (the body's repair and healing mechanisms) and other disturbances (the immune system responds to infections).

In addition to responding to these external influences, the body's system in some cases learns and reconfigures itself (usually to be more efficient) for future response to similar influences. Because of this, the body will respond quite differently the next time the same external influence is experienced (*e.g.,* athletes' hearts and muscles become more efficient, vaccines boost the immune system's response to viruses). This within-person differences in response are known as intra-person variability.

Different people will also typically respond differently to similar influences under similar conditions. In some cases, these differences in response can be quite significant. This is what is known as inter-person variability. The variability of the environment is accounted for because of the way the human dynamical system was defined in chapter 3. Systems that are not well adapted to both types of variability will result in either a failure to meet functional goals or an introduction of significant side-effects.

The fourth reason has to do with the health management system itself. It may exhibit variability because of variability in the behavior of the humans who are part of the health management system, or there may be variability introduced by manufacturing or general variability due to disturbances from the operational environment. These kinds of variability are more controllable than the variability in on the human dynamical system side. Nevertheless, they should be considered.

The failure to meet functional goals and introduction of side-effects are the general mishaps (undesirable outcomes) in the case of health management, and this chapter provides mechanisms for expressing and discussing their acceptable risk (*i.e.,* a model of patient safety), while accounting for variability. In addition two case studies are used to demonstrate the utility of the resulting model of patient safety for developing patient safety criteria (including for subsystems), guiding safety assessments (including for multi-use designs and comparison of equivalent technologies), safety-guided design, discussion of patient safety criteria with design feasibility considerations in mind, exploring impacts of assumptions and rationale on patient safety, and structuring safety arguments.

## 4.2 Health Management Risk and Patient Safety Criteria

The model of health management developed in chapter 3 was centered around individual patients. Because of inter-person variability, the risk (and patient safety criteria) of a health management system must be accounted for at the population level. This can be done in two subtly-different ways, and both approaches are presented.

The ideas are developed in the context of a case study on patients in intensive care units (ICUs) who exhibit what is known as stress-induced hyperglycemia (explained in example 4.1). It is important to note that examples provided from the case study are meant to show what the notions look like in the context of realistic data, and to demonstrate the validity of the notions by showing that they can be mapped to real contexts. They are not meant to demonstrate risk assessments of the particular health management system. In other words, the notions describe how we would like to assess the particular technology and not the process of assessment itself. These notions, and the resulting model of patient safety, can then be used to guide an assessment process, as discussed in section 4.4.2.

**Example 4.1:** Stress-Induced Hyperglycemia and Glycemic Control in the ICU

Surgical and ICU patients have been shown to exhibit what is known as stress-induced hyperglycemia [68], when the body temporarily loses its ability to reduce the blood glucose levels due to the effects of 'stressful' events like surgery or conditions that require admission to an ICU. This results in undesirably-high glucose levels, which as mentioned previously, has both short-term and long-term consequences [111].

Hyperglycemia in the ICU is treated with insulin and the procedure for treatment is described in insulin infusion protocols [98]. Protocols differ in target blood glucose ranges, but in general all protocols have a target range within which the patient is considered in normal condition with respect to blood glucose levels.

### 4.2.1 System Operational Scenario and Outcomes

Before the patient safety criteria is developed, we must define what scenario that the criteria must apply to (*i.e.,* the assumptions we making about the health management scenario), and the outcomes we are interested in given that particular scenario. The scenario is comprised of the nature of population that the health management system is designed for, the general nature of the health management system, and the nature of its interactions with the human dynamical system. The assumption is that inter- and intra-person variability is accounted for in the way the population is defined and the points of variability in the health management system are identified.

Since we are concerned with risk and patient safety issues, the outcomes should reflect the possibility to fail to achieve functional goals and to introduce side-effects. This means there must be at least a functional outcome health metric (which would most likely be metric that is used in decided when to introduce the health management system or a proxy metric related to it), and at least one side-effect outcome metric (which would be developed based on knowledge of the way the health management system interacts with the human dynamical system).

**Example 4.2:** Semi-Automated Glycemia Management in the ICU

Below is the description of the scenario and patient outcomes of interest for the case study.

**Health Management System.** For this case study, the decision logic of the insulin infusion protocol is automated and implemented in software. This software takes in as information the current blood glucose value and makes decisions on how to adjust the insulin infusion rate. A nurse manually takes blood glucose readings, types the information into a computer with the protocol software, and adjusts insulin infusion rate to what the software prescribes. Figure 4.1 shows a conceptual visualization of this scenario (omitting the patient environment).

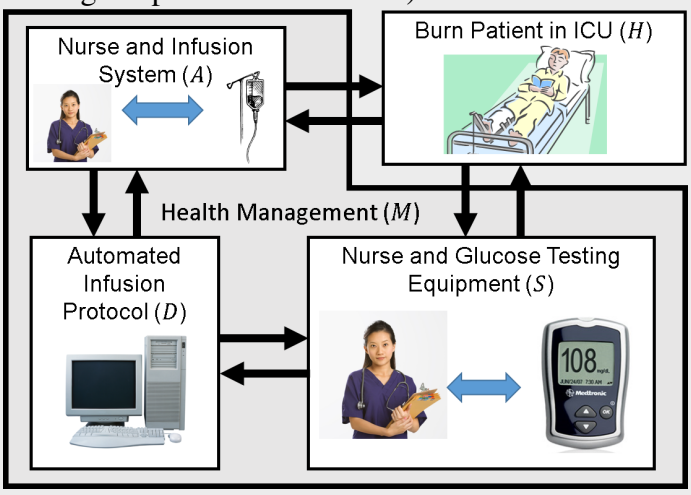



Figure 4.1: The health management system for glycemic control in the ICU.

The main interaction between the health management system and the patient are observing the blood glucose from the patient and infusion of insulin into the patient. This requires physical interactions with both the sensing equipment and the infusion equipment.

*Points of Variability.* For specific protocol setting, the main points of variability in the health management system are the accuracy and timing of sensor information and infusion. In this case study only the variability of sensor information is explored.

**Patient Population.** We assume that the protocol is intended to be used on burn patients in the ICU.

*Inter-Person Variability.* The main inter-person variability considered is the variability in glucose physiology and its response to stress on the body.
*Intra-Person Variability.* The main intra-person variability considered is the specific way in which the stress affects the glucose physiology (and hence the response to insulin infusion) over time.

**Patient Outcomes.** As mentioned above, we must consider both the outcome of the human function we are trying to influence and any side effects that are produced as a consequence of our intervention. For simplicity, we concentrate on a single functional outcome, and a single side-effect outcome.

The function we are interested in is keeping the patient's blood glucose levels within the normal range defined by the hospital infusion protocol. The side-effect we are concerned with is hypoglycemia (low blood glucose levels) caused by an inappropriately high dose of insulin. Note that for this case both outcomes are based on the trajectory of the same state variable, the blood glucose level.

*Functional Health Metric.* One measure of the outcome of an insulin treatment is the percentage of that total of the treatment time that the blood glucose level is within the target range. Since we interested more in undesirable outcomes in safety criteria determinations, we can change the outcome slightly to be the percentage of the time that the blood glucose level is *outside* the target range. The general form of this metric can be represented formally as

$$\mu_{H=\underline{\overline{bg}}} = F_{H=\underline{\overline{bg}}}(x_{bg}(t)) = \frac{|x_{\underline{\overline{bg}}}(t)|}{|x_{bg}(t)|} \tag{4.1}$$

where $x_{bg}(t)$ the sequence of the blood glucose in the time of interest, $x_{\underline{\overline{bg}}}(t)$ is the set of those values in $x_{bg}(t)$ that lie outside the normal range, and $|\cdot|$ indicates the size of the set or sequence.

The visualization of this metric is shown in figure 4.2. The values in the shaded area are the trajectory values that are counted as part of $x_{\underline{\overline{bg}}}(t)$. The target range shown

is 80 to 180 mg/dl, which is the range for a protocol cited by Steil *et al.* as one developed at the University of Washington [98]. The context is the time in which the patient is in the ICU undergoing insulin treatment; hence, $[t_0, t_f]$ will be patient-dependent.

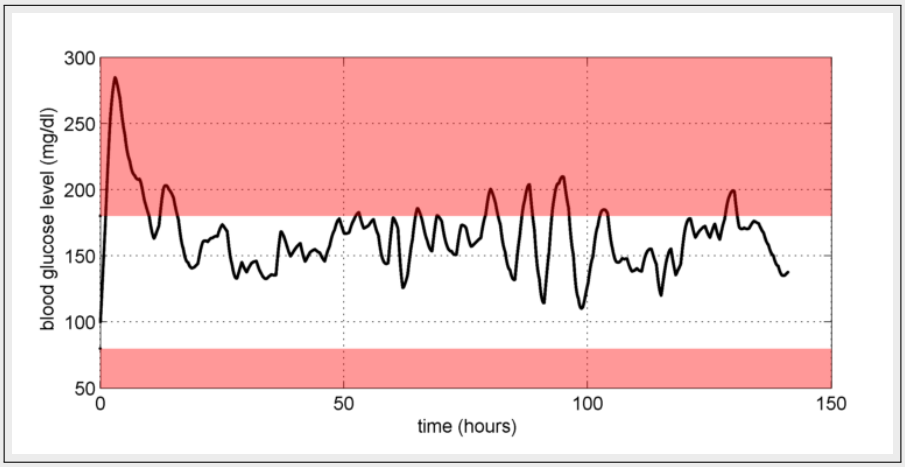



Figure 4.2: Illustration of functional outcome metric for gylcemic control in the ICU. The plot corresponds to the trajectory $x_{bg}(t)$. All values in the red area are counted as part of $x_{\underline{bg}}(t)$.

An alternative way to visualize the health metric is to look at the distribution of blood glucose values of the patient. Since the metric ignores temporal features, it can be interpreted as the amount of the mass of the distribution that is outside the normal range. That is

$$\mu_{H=\underline{bg}} = F_{H=\underline{bg}}(x_{bg}(t)) = 1 - p(x_{\underline{bg}} < x_{bg}(t) < x_{\overline{bg}}) \tag{4.2}$$
$$x_{\underline{bg}} = 80 \text{ mg/dl}$$
$$x_{\overline{bg}} = 180 \text{ mg/dl}$$

This interpretation of the metric is illustrated in figure 4.3. This version of the metrics gives a better picture of what the outcomes for the metric look like.

*Side-Effect Health Metric.* In the protocol, hypoglycemia is defined as blood glucose values below 70 mg/dl. Our side-effect metric is the percent of time spent in the

hypoglycemic range. In the sequence-based approach, this is given by

$$\mu_{H=\underline{bg}_{70}} = F_{H=\underline{bg}_{70}}(x_{bg}(t)) = \frac{|x_{\underline{bg}_{70}}(t)|}{|x_{bg}(t)|} \tag{4.3}$$

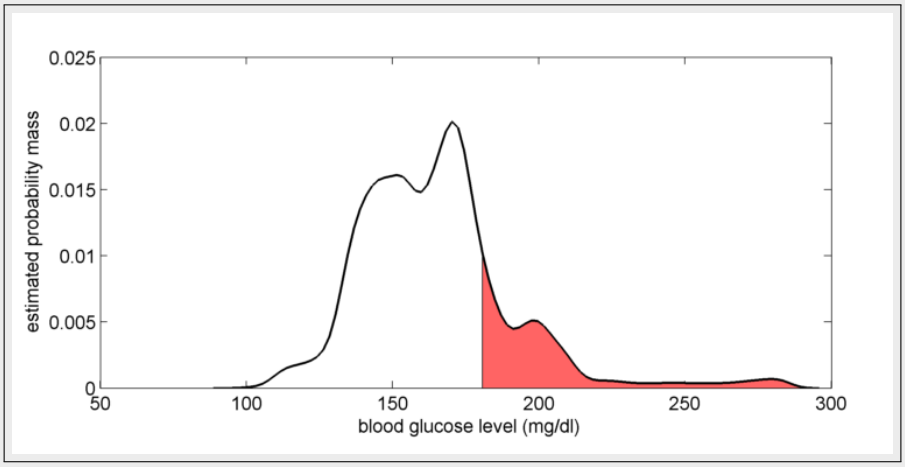



Figure 4.3: Visualization of health metric as a distribution

where $x_{bg}(t)$ the sequence of the blood glucose in the time of interest, $x_{\underline{bg}_{70}}(t)$ is the set of those values in $x_{bg}(t)$ that lie in the hypoglycemic range, and $|\cdot|$ indicates the size of the set or sequence.

In the distribution-based approach, this is given by

$$\mu_{H=\underline{bg}_{70}} = F_{H=\underline{bg}_{70}}(x_{bg}(t)) = p(x_{bg}(t) < 70) \tag{4.4}$$

### 4.2.2 Population-Level Risk

Once the scenario is defined, the next thing is to define the population-level risk metric which helps account for inter-person variability. This metric depends on the idea of a test scenario and a baseline scenario. The test scenario is the health management scenario which includes the technology of concern. The baseline scenario provides a reference with which we can define our notion of risk. It may be a health management scenario, or it could be a scenario where health management has not been introduced. In either case, both scenarios must be able to produce the same kind of health metric outcomes (*i.e.,* the same set of functional and side-effect health metrics).

The population-level risk is a comparison of test scenario health metric outcomes that of the baseline. Because we are interested in inter-person variability, multiple test scenarios must be considered across a representative population of patients where for each scenario, the technology of concern interacts with a different patient from a representative population.

**Notion 3:** Population-Level Risk

The population-level risk metric ($\mu_M$) is a mapping of a set of health metric outcomes ($\{\mu_H\}$) for a 'test' population of patients to a set of risk values, with respect to a set of baseline outcome values ($\{\overline{\mu}_H\}$).

$$\mu_M = F_M(\{\mu_H\}, \{\overline{\mu}_H\}) \tag{4.5}$$

such that

$$\mu_M^i > \mu_M^j \Leftrightarrow \mu_M^i \text{ is higher risk than } \mu_M^j \tag{4.6a}$$
$$\mu_M = 0 \quad \Leftrightarrow \mu_M = F_M(\{\overline{\mu}_H\}, \{\overline{\mu}_H\}) \tag{4.6b}$$

As with other notions, all objects can be multidimensional. Properties 4.6a and 4.6b imply (for each dimension) that 'positive' values denote higher risk and negative values denote lower risk compared to a baseline.

**Notion 4:** Baseline Outcomes

A baseline outcome is a specific outcome value ($\overline{\mu}_H$) of a health metric, chosen to provide a reference for a risk scale in the context of health management. A baseline could be personal, in which case it would be based on the outcome of the health metric on a trajectory for the person outside the context of the specific health management system under consideration. It could also be based on population considerations, in which case it represents some notion of the expected outcome of the metric, which would mean it could be be related to a context similar to or outside the specific health management system under consideration.

An example of a personal baseline could be one related to a person's resting heart rate before being prescribed medication for cardiovascular problems. The idea here is that this

would be compared to some function of that same person's resting heart rate after they have been prescribed these medications. An example of a baseline based on population considerations could be one related to the typical range (or other statistical information) of resting heart rate values for a representative population of people on the medication this person was prescribed.

The use of baselines (personal or otherwise) acknowledges that in many cases, even though health management may improve health outcomes (*i.e.,* move patients dynamics further away from unacceptable modes and closer to an acceptable mode), it may not always result in dynamics that would be considered acceptable outside the context of health management (*i.e.,* in the case where the body maintains its own health).

There are two subtly-different ways in which population-level risk metrics can be defined. The main difference is in the way baselines are used in the definition. Both are described below.

### Direct Population-Level Risk

This approach is a direct application of the general population-level risk definition given in notion 3. It is analogous what is done in clinical trials with a control group and an experimental group which are two physically different populations.

**Notion 5:** Direct Population-Level Risk Metric

The direct population-level risk metric is a mapping of a set of test individual health metric and a set of baseline health metrics to a set of values suitable for comparison.

$$\mu_M = F_M(\{\mu_H^{H_1}, \ldots, \mu_H^{H_n}\}, \{\mu_H^{\overline{H}_1}, \ldots, \mu_H^{\overline{H}_m}\}) \tag{4.7}$$

where $\{\mu_H^{H_1}, \ldots, \mu_H^{H_n}\}$ is the set of test health metrics and $\{\mu_H^{\overline{H}_1}, \ldots, \mu_H^{\overline{H}_m}\}$ is the set of baseline health metrics. The (range of) health management scenario(s) is expected to be the same for each patient in the test population.

Note that this approach requires identifying a set of baseline health metric outcomes to compare to. A concrete example of this approach is given below.

**Example 4.3:** Direct Population-Level Risk of Glycemia Control in the ICU

Here we compare a set of health metric values ($\{\mu_H^{H_1}, \ldots, \mu_H^{H_n}\}$) from the test population to the set of health metric values from a baseline population ($\{\mu_H^{\overline{H}_1}, \ldots, \mu_H^{\overline{H}_m}\}$), where the health metric outcomes are obtained using the health metric definitions in equations 4.1 or 4.2 and 4.3 or 4.4. The baseline scenario here could be healthy patients or other hospitalized patients but not in an ICU.

In this case, for each metric (functional and side-effect), we interested in two things. First, is difference between the average value from the test population and that from the baseline population.

$$\mu_{M,1} = \begin{bmatrix} \mu_{M,1=\overline{bg}} \\ \mu_{M,1=\underline{bg}_{70}} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \cdot \sum_{i=1}^{n} \mu_{H=\overline{bg}}^{H_i} \\ \frac{1}{n} \cdot \sum_{i=1}^{n} \mu_{H=\underline{bg}_{70}}^{H_i} \end{bmatrix} - \begin{bmatrix} \frac{1}{m} \cdot \sum_{i=1}^{m} \mu_{H=\overline{bg}}^{H_i^*} \\ \frac{1}{m} \cdot \sum_{i=1}^{m} \mu_{H=\underline{bg}_{70}}^{H_i^*} \end{bmatrix} \tag{4.8}$$

This is illustrated in Figure 4.4.

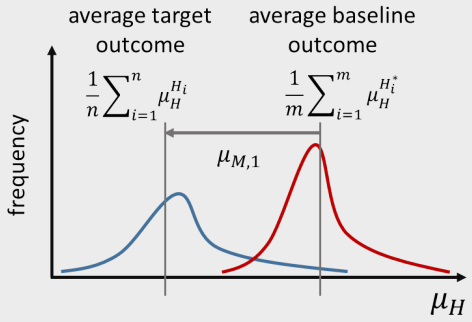



Figure 4.4: Illustration of population-level risk metric for Artificial Pancreas considering the average of outcomes between the target scenario and the baseline.

Second, for the outcome value $\mu_H^{*0.75}$ such that 75% of the baseline have outcomes worse than this value, we want to know how much of the test population have

outcomes worse than this value.

$$\mu_{M,2} = \begin{bmatrix} p\left(\mu_{H=\underline{\overline{bg}}} > \mu^{*0.75}_{H=\underline{\overline{bg}}}\right) \\ p\left(\mu_{H=\underline{\overline{bg}}_{70}} > \mu^{*0.75}_{H=\underline{\overline{bg}}_{70}}\right) \end{bmatrix} \tag{4.9}$$

This is illustrated in Figure 4.5

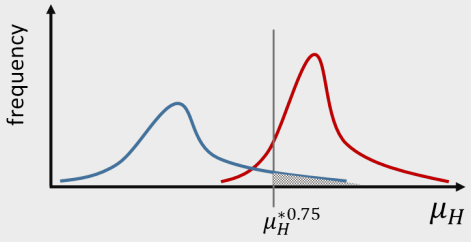



Figure 4.5: Illustration of population-level risk metric for Artificial Pancreas considering the overlap of outcomes between the target scenario and the baseline.

The population-level risk metric in this case is a four-dimensional object.

### Individual-Risk-Based Population-Level Risk

This approach relies on the notion of an individual baseline in order to compute an individual risk metric. The population-level risk metric then relies on a collection of individual risk metrics from a representative population.

**Notion 6:** Individual Risk Metric

The individual risk metric ($\mu_R$) is a mapping of a set of health metric outcomes for an individual patient in the context of interactions with a health management system to a set of risk values, with respect to the respective individual baseline outcome values. ($\overline{\mu}_H$)

$$\mu_R = F_R(\mu_H, \overline{\mu}_H) \tag{4.10}$$

such that

$$\mu_R^i > \mu_R^j \Leftrightarrow \mu_R^i \text{ is higher risk than } \mu_R^j \tag{4.11a}$$
$$\mu_R = 0 \quad \Leftrightarrow \mu_R = F_R(\overline{\mu}_H, \overline{\mu}_H) \tag{4.11b}$$

As with other notions, all objects can be multidimensional. Properties 4.11a and 4.11b imply (for each dimension) that 'positive' values denote higher risk and negative values denote lower risk compared to a baseline. Computing an individual risk metric is analogous to computing a population-level risk metric for an individual patient

A concrete example of the individual risk metric is given below.

**Example 4.4:** Individual Risk of Semi-Automated Glycemia Management in the ICU

Since we have two health metrics, we will end up with two risk metrics (one for the functional outcome and the other for the side-effect). This results in a two dimensional risk metric ($\mu_R = [\mu_{R=\overline{bg}}, \mu_{R=\underline{bg}_{70}}]$) which consists of the functional risk metric ($\mu_{R=\overline{bg}}$) and the side-effect risk metric ($\mu_{R=\underline{bg}_{70}}$). The particular dimension each one occupies is inconsequential, but the convention we will adopt is to put the function metric in the first dimension. Though we could find a way to combine these two risk metrics into one, working with the two-dimensional metric is more informative since we can see more easily whether the issue is a functional issue or side-effect issue.

**Functional Risk Metric** Since our functional metric outcome (percentage time outside the normal range) is a scalar value, our function metric is the difference between the functional health metric and an individual (not necessarily personal) baseline ($\overline{\mu}_{H=\overline{bg}}$).

$$\mu_{R=\overline{bg}} = \mu_{H=\overline{bg}} - \overline{\mu}_{H=\overline{bg}} \tag{4.12}$$

Since the health metric ranges in values from 0 to 1 (or 0% to 100%), the risk metric will range from $-\overline{\mu}_{H=\overline{bg}}$ to $1 - \overline{\mu}_{H=\overline{bg}}$.

**Side Effect Risk Metric.** The risk metric for side effect outcome is similar to the that of the functional outcome

$$\mu_{R=\underline{bg}_{70}} = \mu_{H=\underline{bg}_{70}} - \overline{\mu}_{H=\underline{bg}_{70}} \tag{4.13}$$

where $\overline{\mu}_{H=\underline{bg}_{70}}$ is the individual baseline.

Having defined the individual risk metric, we can now define the population level risk metric, which is our primary concern.

**Notion 7:** Individual-Based Population-Level Risk Metric

The individual-risk-based population-level risk metric is a mapping of a set of individual risk metrics from a population to a set of values suitable for comparison.

$$\mu_M = F_M(\{\mu_R^{H_1}, \ldots, \mu_R^{H_n}\}) \tag{4.14}$$

where the $\mu_R^{H_i}$'s are individual risk value. As with other notions, all objects can be multidimensional.

Note that this form of the population-level risk is a bona fide risk metric as it is a function of 'test' population health metric outcomes and baseline outcomes. Each individual risk ($\mu_R^{H_i}$) is a function of 'test' outcomes and baseline outcomes (*i.e.,* $\mu_R^{H_i} = F_R(\mu_H^{H_i}, \overline{\mu}_H^{H_i})$ ), and this individual-based population-level risk metric is a function of the output this individual risk function. By composition of functions, this makes the individual-based population risk a function of the inputs to the individual risk metric (*i.e.,* $F_M(F_R(\mu_H^{H_i}, \overline{\mu}_H^{H_i}))$ makes $F_M$ a function of $(\mu_H^{H_i}, \overline{\mu}_H^{H_i})$), and these inputs are the kind of inputs a population-risk metric as defined in notion 3 requires.

A concrete example of this kind of population-level risk metric is given below.

**Example 4.5:** Individual-Basd Population-Level Risk of Glycemica Control in the ICU

A simple risk metric we may be interested in is what proportion of the population does worse than the baseline (*i.e.,* have individual risk values greater than 0) for each individual risk metric. For each dimension of the risk metric, this can be viewed as a

proportion

$$\mu_M = F_M(\{\mu_R^{H_1}, \ldots, \mu_R^{H_n}\}) = \frac{|\{\mu_R^{H_i} | \mu_R^{H_i} > 0\}|}{|\{\mu_R^{H_1}, \ldots, \mu_R^{H_n}\}|} \tag{4.15}$$

where the numerator is the size of the set of values that do worse than the baseline and the denominator is the size of the population (or the total number of individual risk metrics).

Figure 4.6 shows the visualization of this metric. At the top right is the value of the population risk metric based on equation 4.15 applied to each dimension of the individual risk metric outcomes. At the top left is the distribution of individual side-effect risk values with the proportion highlighted in red showing corresponding to the value of the vertical dimension of the population-level risk. At the bottom right is the distribution of individual functional risk values with the proportion highlighted in red corresponding to the value of horizontal dimension of the population-level risk.

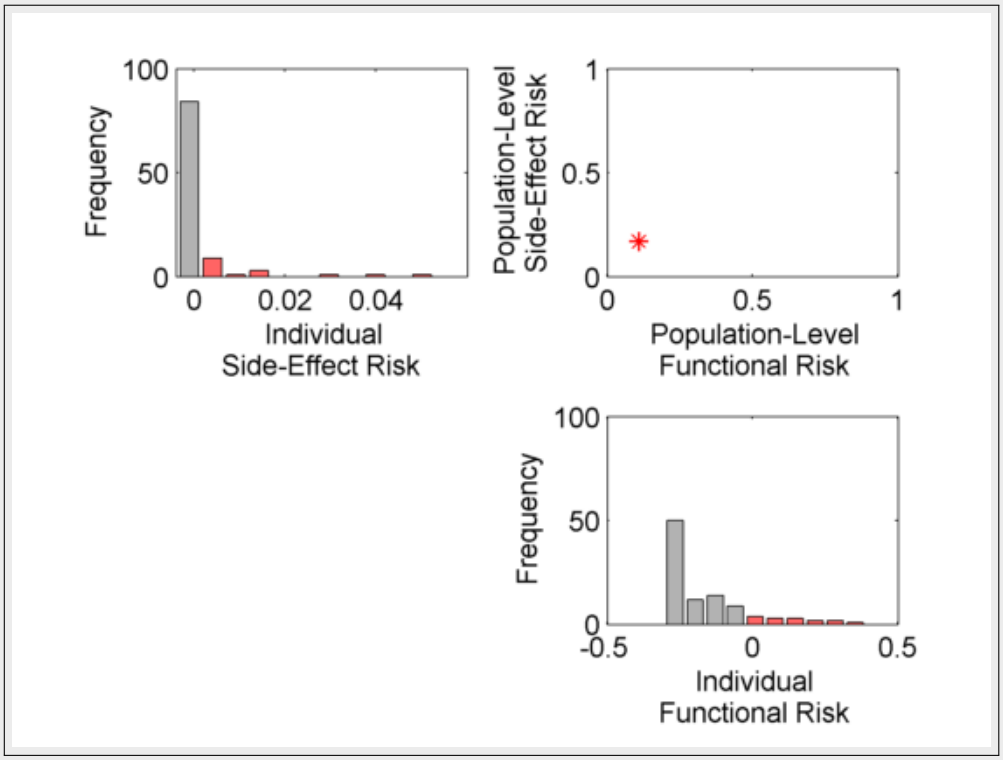



Figure 4.6: Visualization of the two-dimensional risk metric.

**A Note on Adaptive or Personalized Health Management**

It is important to note the approaches to defining population-level risk metric account for personalized and adaptive health management systems. Each approach assumes that the health metric outcomes of interest are the result of patient interactions with the specified health management system, hence if this system is adaptive or customized in some way, then it would be accounted for in the system operational scenario definition (addressed in section 4.2.1). What would remain consistent across the population would most probably be the method of personalization or adaptation. Hence, the population risk metric would reflect the risk of the method of personalization or adaptation across the population.

### 4.2.3 Acceptable Population-Level Risk

The final piece of the model is the acceptable population risk. It is a set of criteria that describes what population-level risk values are considered acceptable, and in effect, which health management scenarios are considered safe.

**Notion 8:** Acceptable Population-Level Risk Criteria

The acceptable risk criteria ($\overline{\mu}_M$) defines a set of conditions that the population-level risk metric must satisfy in order for the population-level outcomes to be considered acceptable.

**Example 4.6:** Acceptable Population-Level Risk for Semi-Automated Glycemia Management in the ICU

In this example, we use the individual-risk-based population metric. A criteria may be to limit the proportion of the population that does worse than the baseline for each dimension of the risk metric.

$$\overline{\mu}_M : \mu_{M=\underline{\overline{bg}}} < \overline{\mu}_{M=\underline{\overline{bg}}} \text{ and } \mu_{M=\underline{bg}_{70}} < \overline{\mu}_{M=\underline{bg}_{70}} \tag{4.16}$$

The creates a region (as shown in Figure 4.7) where any point in that region is acceptable and hence satisfies the criteria. The point shown is unacceptable.

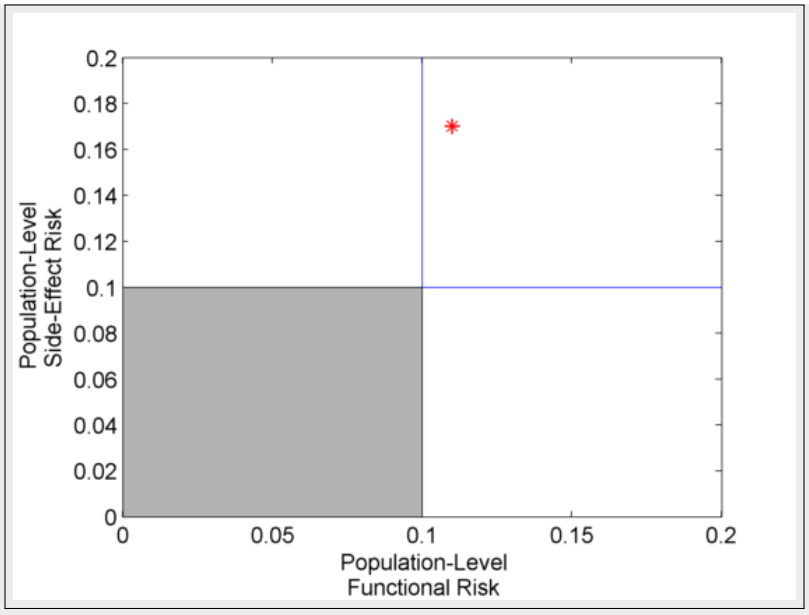



Figure 4.7: Visualization of the two-dimensional risk metric with the acceptable risk region.

## 4.3 The Model of Patient Safety

We can now put the ideas developed in chapter 3 and above together to form the model of patient safety. The conceptual diagram is shown in figure 4.8. One way to interpret the model is as follows. Any patient and the technology designed to help them exist together (in an environment) in what is called a health management scenario ($M \leftrightarrow H$). The health management scenario is designed to improve the health metrics of the patient in order to reduce the 'natural' unacceptable risks (as measured by the health metrics) to acceptable levels. A health management scenario is considered safe for a patient if a convincing argument can be made that it brings overall risk (including the newer ones introduced by having the system) to acceptable levels.

The patient safety model here models the risk of having the health management scenario compared to a baseline situation without the specific health management scenario under consideration. The baseline then represents a risk we are willing to accept. To account

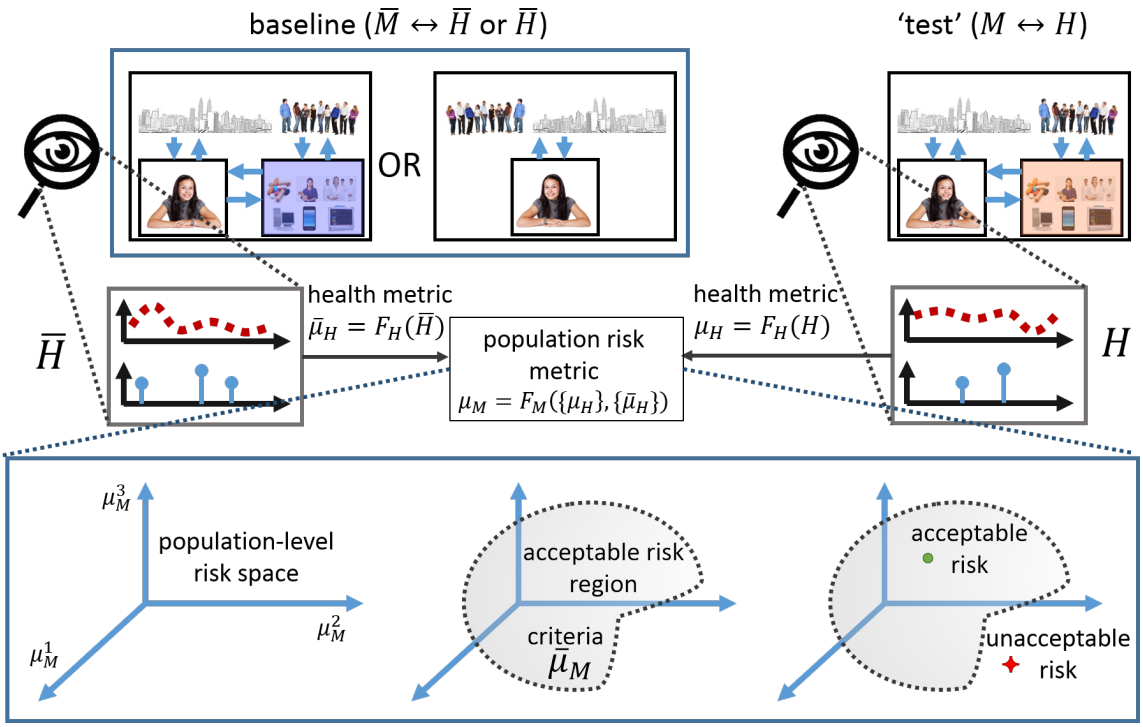



Figure 4.8: Conceptual diagram of the patient safety model

for the inevitable variability in the health management scenario due to variability in the patients and components, a population-level risk metric ($\mu_M$) is used, which takes as input two sets of health metrics ('test' and baseline) from a representative population.

Population-level risk metric defines a risk space, an n-dimensional space accounting for risk related to the different health metrics under consideration (figure 4.8 shows a 3-dimensional space). To define the population level risk one needs a baseline set of health metrics ($\{\overline{\mu}_H\}$) and a 'test' set of health metrics ($\{\mu_H\}$). The risk space is therefore the possible range of values of the population risk metric given a specific baseline set of values and the possible range of values the health metrics in general.

The baseline set of metrics act as a reference for the risk space. Hence the $\mathbf{0}$ vector in the risk space represents the output of the risk metric applied to the baseline, and for any dimension, positive values denote health management risks worse than the baseline and negative values denote outcomes better than the baseline.

For a given risk space, the acceptable risk criteria ($\overline{\mu}_M$) defines a portion of the space as the acceptable risk region. Applying the risk metric to a 'test' set of health metrics from

a specific health management scenario and the baseline set of values used to define the particular risk space will produce a point in the risk space, telling us the risk of the scenario relative to the baseline. If this point lies in the acceptable risk region, then the risk of that scenario is considered acceptable, and if the point lies outside this region, the risk of that scenario is considered unacceptable.

The health metrics considered must include at least one functional outcome health metric (which is related to the health metric used to determine that health management must be introduced), and a side-effect outcome metric. These account for the fact that variability could cause the health management system to fail to achieve its intended function (which in health is a patient safety issue) or can introduce unwanted unacceptable dynamics in the person.

In an abuse of terminology, the model described above is in some sense denotational. This makes it more of a metamodel since it encompasses the class of models for patient safety for specific scenarios. The next section demonstrates how this model can be operationalized for different purposes related to reasoning about the safety of medical technologies.

## 4.4 Utility of the Model

This section illustrates some of uses of this model in reasoning about the safety of medical technologies. The uses considered are development of patient safety criteria, assessment technologies against a specific criteria, safety-guided design of a (part of) a particular technology, discussions of safety criteria, and exploration of the impact of assumptions and rationale on safety criteria and acceptable risk. There is also a brief discussion on its use in structuring safety arguments.

Two case studies are used demonstrate the utility of the model for these purposes: the one on glycemia management in the ICU already introduced above, and one related to the artificial pancreas [29], introduced below. Both are related to management of blood glucose levels but in a different health management contexts. It is important to note that the case studies are only illustrative and any results have no bearing on the real-world versions of these systems.

Also, the order of presentation does not represent the order of use of the model in a design process. Rather, the subsections are organized so that subsequent subsections can build on ideas introduced in the previous ones. For example, technologies are assessed

against a specific criteria so illustrating the development of criteria first helps make the illustration of assessment of technologies easier.

**Example 4.7:** Glycemia Management with the Artifical Pancreas

The artificial pancreas [29] is a system for managing blood glucose levels in a 'continuous' manner. It consists of a sensor for inferring blood glucose, a pump for infusing insulin, and software for deciding on insulin infusion rates based on information from the sensor and other sources. The sensor provides inferences of blood glucose values every five minutes and hence control decisions are made at this rate as well. The general conceptual picture of this health management setup is shown in figure 4.9.

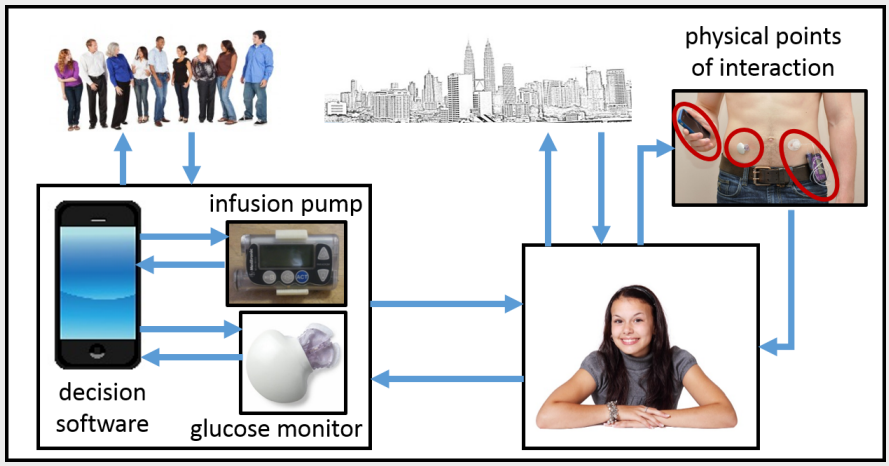



Figure 4.9: The conceptual setup of the artificial pancreas system

There are a number of differences between this health management scenario and the ICU infusion protocol scenario described previously. First is that the artificial pancreas is targeted at Type I diabetics whose ability to bring blood glucose levels down have been permanently impaired, whereas the ICU system is targeted at a population where this ability is temporarily impaired. Second, the artificial pancreas operates at a finer time resolution (an order of magnitude smaller) than the ICU system. Lastly, the communication between the sensing, decision-making, and actuation technologies in the artificial pancreas is automated, whereas a nurse must serve as the communication medium in the ICU system.

### 4.4.1 Patient Safety Criteria Development

One of the questions in reasoning about safety of medical technologies is under what conditions do we consider a particular health management strategy acceptable? This corresponds to developing the parts of the patient safety model shown in the conceptual diagram in Figure 4.10. The main goal of patient safety criteria is to develop the risk space and identify the acceptable risk region of that space.

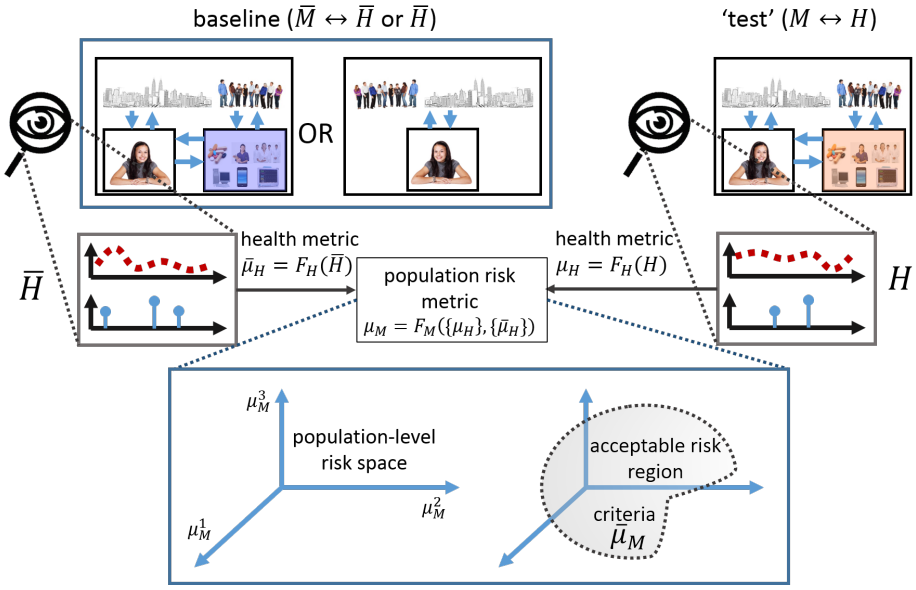



Figure 4.10: Conceptual depiction of patient safety criteria development. The main goals are to establish the population-level risk space and the acceptable risk region.

A manufacturer may develop the criteria and argue to the FDA that the criteria is valid; the FDA may specify the criteria for manufacturer to follow; health practitioners may provide input to either or both; or all three may jointly develop the criteria. Regardless of who defines it, the criteria consists of the following.

**The Health Management Scenario ($M \leftrightarrow H$).** The criteria must state clearly which health management scenario it applies to. We must define as clearly as possible the population of concern, the goals of the specific health management system, and the ways in which it interacts with a patient. The level of variability in all pieces to be considered can also be defined.

**The Health Metrics ($\mu_H$).** Related to the health management scenario are the health metrics. We must define how the functional outcome (what the health management system is helping with) is measured, as well as the potential side-effects as a result of the introduction of the health management system and how are these also measured (in terms of health metrics).

**The Population-Level Risk Metric ($\mu_M$).** It is expected that the system would account for and be robust to inter-person variability as much as possible. We must define what we mean by the risk of the system with respect to a baseline with this variability in mind, either using an individual-risk-based approach or a direct population level approach. We need not specify the particular baseline values, though we could, but we must assume that there will be a baseline.

**The Acceptable Risk Level ($\overline{\mu}_M$).** Eventually, we must define the level of population risk we deem acceptable. This is should be based on the outcome (values) of the population risk metric as applied to the case of the particular health management system.

**Example 4.8:** Patient Safety Criteria for Semi-Automated Glycemia Management in the ICU

The criteria for this scenario was developed as we introduced the various notions. Below is a summary what the various pieces were.

**Health Management Scenario.** The population of concern are burn patients (adolescent or adult) in the ICU suffering from stress-induced hyperglycemia. The system is an insulin infusion protocol that adjusts insulin infusion rates once every hour based on blood glucose readings at the time of adjustment. Here, we are concerned with the general nature of the system and its behavior (*i.e.,* the fact that it records glucose and adjust insulin every hour). Later, in the assessment, we would provide details on the health management components.

**Health Metrics.** The functional outcome metric ($\mu_{H=\overline{bg}}$) is defined as the percentage of the total time on the protocol that the patient's $\overline{\text{blood}}$ glucose is outside the target range of 80 to 180 mg/dl, using the same protocol previously mentioned that

Steil *et al.* cite as one developed at the University of Washington [98]. The context is the time in which the patient is in the ICU undergoing treatment, making $[t_0, t_f]$ patient-dependent.

The side-effect metric ($\mu_{H=\underline{bg}_{70}}$) is the percentage of the total time on the protocol that the patient's blood glucose is in the hypogylcemic range (below 70 mg/dl). This is results in a two-dimensional health metric consisting of the functional outcome metric and the side-effect metric. Using a distribution of blood glucose values approach, the functional outcome metric is

$$\begin{aligned} \mu_{H=\overline{\underline{bg}}} = F_{H=\overline{\underline{bg}}}(x_{bg}(t)) = 1 - p(x_{\underline{bg}} < x_{bg}(t) < x_{\overline{bg}}) \\ x_{\underline{bg}} = 80 \text{ mg/dl} \\ x_{\overline{bg}} = 180 \text{ mg/dl} \end{aligned} \tag{4.17}$$

and the side-effect metric is

$$\mu_{H=\underline{bg}_{70}} = F_{H=\underline{bg}_{70}}(x_{bg}(t)) = p(x_{bg}(t) < 70) \tag{4.18}$$

where $x_{bg}(t)$ is the blood glucose trajectory and $p(\cdot)$ can be interpreted as the ratio of blood glucose values that satisfy the condition in the parenthesis to the total number of blood glucose values in the trajectory. The time of interest is the total time the patient interacts with the protocol.

**Population-Level Risk.** This is based on the individual risk approach. The individual risk for each patient consists on a functional risk and side-effect risk. For each, there is a baseline health metric outcome ($\overline{\mu}_{H=\overline{\underline{bg}}} = 30\%$ for the functional and $\overline{\mu}_{H=\overline{\underline{bg}}} = 0\%$ for the side-effect). The individual risk is the actual outcome (percentage outside or in the respective range) minus the baseline outcome.

$$\mu_R = \begin{bmatrix} \mu_{R=\overline{\underline{bg}}} \\ \mu_{R=\underline{bg}_{70}} \end{bmatrix} = \begin{bmatrix} \mu_{H=\overline{\underline{bg}}} \\ \mu_{H=\underline{bg}_{70}} \end{bmatrix} - \begin{bmatrix} \overline{\mu}_{H=\overline{\underline{bg}}} \\ \overline{\mu}_{H=\underline{bg}_{70}} \end{bmatrix} \tag{4.19}$$

The population-level risk (for an appropriately-chosen representative population) is then a two-dimensional vector representing the proportion of the population that

does worse than the baseline on each outcome

$$\mu_M = \begin{bmatrix} \mu_{M=\overline{\underline{bg}}} \\ \mu_{M=\underline{bg}_{70}} \end{bmatrix} \begin{bmatrix} p(\mu_{R=\overline{\underline{bg}}} > 0) \\ p(\mu_{R=\underline{bg}_{70}} > 0) \end{bmatrix} \tag{4.20}$$

**Acceptable Population-Level Risk.** In this case we put a limit on the population level risk value, where in order for the risk to be acceptable no more than 10% of the population should do worse than the baseline on each dimension.

$$\overline{\mu}_M : \mu_{M=\overline{\underline{bg}}} < 0.1 \wedge \mu_{M=\underline{bg}_{70}} < 0.1 \tag{4.21}$$

**Example 4.9:** Patient Safety Criteria for Artificial Pancreas

We can develop a criteria for this case as well. In this case, we use the direct population-level approach to defined the risk.

**Health Management Scenario.** The population of concern are Type I diabetics (adolescents or adult). The system is the artificial pancreas system described in example 4.7 where the decision to adjust insulin is made every five minutes based on blood glucose and other information (including meals).

**Health Metrics.** The functional outcome metric ($\mu_{H=\overline{\underline{bg}}}$) is similar to the ICU case and is the defined as the percentage of the total time that the patient's blood glucose is outside the normal range, but the range is now 70 to 180 mg/dl, which is what is typically used in artificial pancreas considerations [81, 64].

The side-effect metric ($\mu_{H=\underline{bg}_{70}}$) is the percentage of the total time on the protocol that the patient's blood glucose is in the hypogylcemic range (below 70 mg/dl). This is results in a two-dimensional health metric consisting of the functional outcome metric and the side-effect metric. Using a distribution of blood glucose values approach, the

functional outcome metric is

$$\begin{aligned} \mu_{H=\overline{\underline{bg}}} = F_{H=\overline{\underline{bg}}}(x_{bg}(t)) = 1 - p(x_{\underline{bg}} < x_{bg}(t) < x_{\overline{bg}}) \\ x_{\underline{bg}} = 70 \text{ mg/dl} \\ x_{\overline{bg}} = 180 \text{ mg/dl} \end{aligned} \tag{4.22}$$

and the side-effect metric is

$$\mu_{H=\underline{bg}_{70}} = F_{H=\underline{bg}_{70}}(x_{bg}(t)) = p(x_{bg}(t) < 70) \tag{4.23}$$

where $x_{bg}(t)$ is the blood glucose trajectory and $p(\cdot)$ can be interpreted as the ratio of blood glucose values that satisfy the condition in the parenthesis to the total number of blood glucose values in the trajectory. The time of interest is the total time the patient interacts with the protocol.

**Population-Level Risk.** This is based on the direct population-level approach. Here we compare a set of health metric values ($\{\mu_H^{H_1}, \ldots, \mu_H^{H_n}\}$) from the target scenario to the set of health metric values from a baseline scenario ($\{\mu_H^{H_1^*}, \ldots, \mu_H^{H_m^*}\}$). The baseline scenario here is diabetics who manage the blood glucose with a sensor and pump but without the aid of the artificial pancreas decision-making platform.

In this case, for each metric (functional and side-effect), we interested in two things. First, is difference between the average value from the target scenario and that from the baseline scenario

$$\mu_{M,1} = \begin{bmatrix} \mu_{M,1=\overline{\underline{bg}}} \\ \mu_{M,1=\underline{bg}_{70}} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \cdot \sum_{i=1}^{n} \mu_{H=\overline{\underline{bg}}}^{H_i} \\ \frac{1}{n} \cdot \sum_{i=1}^{n} \mu_{H=\underline{bg}_{70}}^{H_i} \end{bmatrix} - \begin{bmatrix} \frac{1}{m} \cdot \sum_{i=1}^{m} \mu_{H=\overline{\underline{bg}}}^{H_i^*} \\ \frac{1}{m} \cdot \sum_{i=1}^{m} \mu_{H=\underline{bg}_{70}}^{H_i^*} \end{bmatrix} \tag{4.24}$$

This is illustrated in figure 4.11

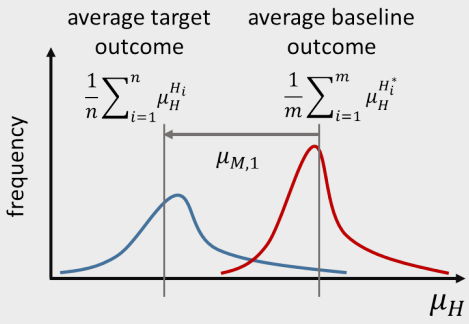



Figure 4.11: Illustration of population-level risk metric for Artificial Pancreas considering the average of outcomes between the target scenario and the baseline.

The second is how much of the population from the target scenario overlaps with the worst 75% of the baseline population in terms of outcomes

$$\mu_{M,2} = \begin{bmatrix} p\left(\mu_{H=\underline{bg}} > \mu_{H=\underline{bg}}^{*0.75}\right) \\ p\left(\mu_{H=\underline{bg}_{70}} > \mu_{H=\underline{bg}_{70}}^{*0.75}\right) \end{bmatrix} \tag{4.25}$$

Here, $\mu_H^{*0.75}$ the health metric outcome value such 75% of the baseline population have a health metric outcome value worse (greater) than this. This is illustrated in figure 4.12

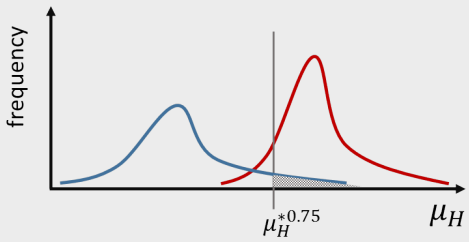



Figure 4.12: Illustration of population-level risk metric for Artificial Pancreas considering the overlap of outcomes between the target scenario and the baseline.

The population-level risk metric in this case is a four-dimensional object.

**Acceptable Population-Level Risk.** In this case, we put a limit on each of the four dimensions of the population-level risk metric.

For $\mu_{M,1}$, where we are interested in the difference in average outcomes, a negative value indicates that the average outcome in the target scenario is better than the average outcome in the baseline scenario. In the functional metric case, our criteria is that the average target outcome be 50% better than the case for the baseline, and for the side-effect case we require that it just be better. These would be expressed

$$\overline{\mu}_{M,1} : \mu_{M,1} < \begin{bmatrix} -\left(0.5 \cdot \frac{1}{m} \cdot \sum_{i=1}^{m} \mu_{H=\overline{bg}}^{H_i^*}\right) \\ 0 \end{bmatrix} \tag{4.26}$$

In terms of the overlap,for the functional metric, we want no more than 20% of the population to be in this overlap range, and in the side-effect case we want no more than 30% to be in this overlap range. This would be expressed as

$$\overline{\mu}_{M,2} : \mu_{M,2} < \begin{bmatrix} 0.2 \\ 0.3 \end{bmatrix} \tag{4.27}$$

From the above, it is clear that the safety criteria specifies the information one must gather and the computations to undertake in order to find out the population-level risk (as defined by the criteria) and whether this risk value is acceptable or not (also as defined by the criteria).

Note that the criteria developed above is quite general. For example, the specific characteristics of the patient population that would be considered appropriate is not explicitly stated. Neither is how the variability in the different components should be explored. These can definitely be specified if whoever is developing the criteria wishes to do so. Leaving it unspecified only means that the choices are made when gathering information for assessing a technology against the criteria (as shown in section 4.4.2).

### Subsystem Patient Safety Criteria Development

The above criteria development approach made no reference to a specific part of the health management system. It is implies that a manufacturer has complete design control of every part of the health management system. This is often not the case: a manufacturer is usually

responsible for a part of the overall system. In this case, we need to develop patient safety criteria for pieces of the health management system.

No new notions need to be added here in order to do this, but the way the health management scenario is described must change. We must now clearly state the health management subsystem (medical technology) under consideration. If there is direct (physical) interactions between this subsystem and the patient, potential side-effects from this interaction must be considered.

The conceptual diagram for highlighting a subsystem is shown in figure 4.13. The rest of the model is omitted and the health management scenario portions are highlighted. Notice that the subsystem is only highlighted in the test scenario since this is the technology we are interested in.

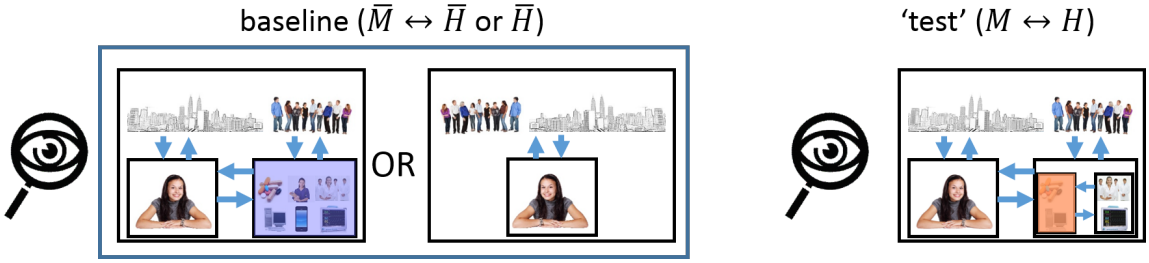



Figure 4.13: Conceptual diagram of a health management scenario description highlighting the subsystem in the test scenario.

In our example case studies, the metrics part of the criteria remain the same. For the semi-automated glycemia management case, if we are interested in the decision-making software, then that part of the system would be highlighted as what the criteria is being developed for. In the artificial pancreas case, we could be interested in any of the three pieces, in which case the specific piece would be highlighted.

Note that the fact that we are interested in a particular subsystem does not change the overall patient safety criteria for the particular health management system. What considering a subsystem does is define an interface between that subsystem and the rest of the health management system. In this case, we have to be explicit about the assumptions made about the parts of the health management system not under consideration. The use of an interface means that the parts of the system not under consideration need to be described in as black-box a manner as possible, especially if there are different possible instantiations of those parts.

The assumptions about the behaviors of other parts not under consideration can be used in two ways. First is that a manufacturer can state it as the expected behaviors that other parts must conform to should they all be used in the same health management system. If the subsystem under consideration is deemed acceptable, and the other parts do conform to the interface behaviors in operation, then we can assume that to be acceptable operation. Second is that the FDA can specify the behaviors of the other parts as behaviors that the manufacturer must assume and design for. The subsystem under consideration in this case is deemed acceptable if it produces an acceptable population risk given the assumptions.

The rest of the chapter will concentrate on the realistic case of designing and reasoning about a subsystem of the health management system. The term health management scenario will still refer to the whole health management system, but the part of primary interest will be the subsystem being considered.

### 4.4.2 Guided Assessment of Patient Safety of a Specific Technology

An assessment is the determination of the population-level risk of a specific health management scenario according to a given criteria. We may be interested in this risk value or we may be interested in a binary decision on whether this risk is acceptable (meets the patient safety criteria) or not. Both cases are illustrated in the concept diagram in figure 4.14. The values shown are two possible values for the same health management system (*i.e.,* the system may result in one or the other but not both).

Note that in the assessment, the specific choices made for both the baseline and test health management scenarios (patient population, behaviors of parts of the health management system, and interactions) must be explicitly stated. However, the since the assessment is not necessarily concerned with why the design results in the specific outcomes, details of the internals of the design under consideration are not necessary.

The assessment is called ‘guided’ because the model provides a guide for what information must be gathered and what factors must generally be considered, but user of the model must make some judgments on the specifics of what are included in the assessment. Depending on where in the design process an assessment is made, the information used could come from simulation models or from actual field data. Lastly, if the assessment scenario involves one party reviewing the information provided by another, the assessor can question the choices made to produce the outcome. These choices include the amount of variability considered and the choice of representative populations (both baseline and test populations).

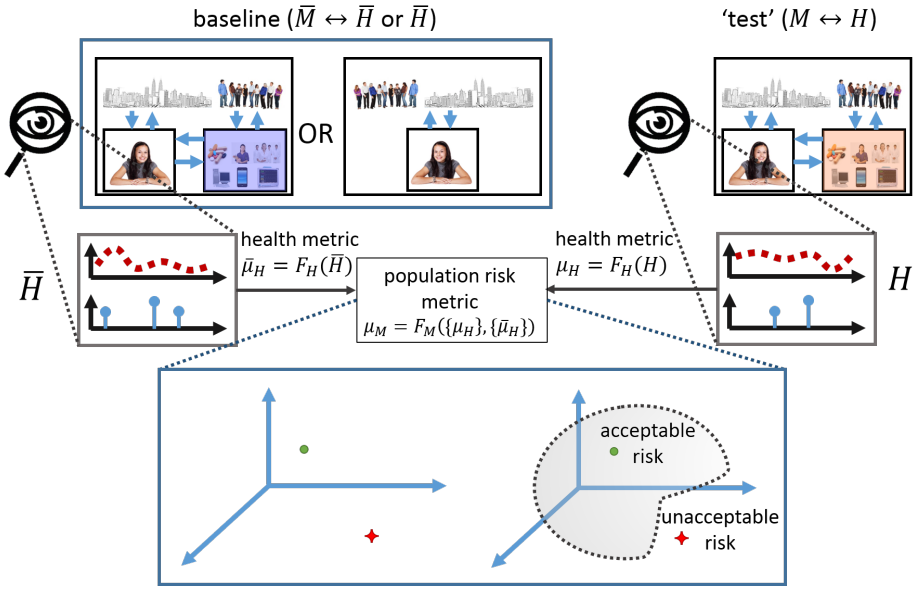



Figure 4.14: Conceptual depiction of assessment of a particular design. The diagram indicates that the outcome of the assessment could be the population-level risk according to a given criteria (which could be one of the two values shown), or a binary decision on whether this risk (and hence design) is acceptable (meets the criteria) or not.

**Example 4.10:** Assessment of Patient Safety of Infusion Decision-Making Software for Glycemia Management in the ICU

Here we are assessing the safety of only the decision-making software. The intrinsic behavior of the software remains the same across the patient population. This example assumes, for simplicity that the main safety concerns are with the results of insulin infusion (*i.e.,* it does not consider issues like risks associated with obtaining the blood glucose sample).

The data used comes from simulations, but similar information could have also been obtained from field tests. (The simulations were run by Scott Popplewell and his team working on a project with at Edward Ortiz at the University of Virginia Center for Diabetes Technologies, and was based on a simulator developed for work on optimizing ICU protocols [82]. I only reused the blood glucose traces from the original simulations.)

**Health Management Scenario Considered.** The health management system consists of insulin infusion decision-making software (which is what we are interested in assessing), nursing staff, a blood glucose meter, and the infusion system. The nurse obtains the blood glucose readings and types them to the decision-making software, and also adjusts the infusion rate to that output by the software.

100 patients were considered in this assessment, and nothing was done to restrict the general behavior of the patients. This means the 100 patients produce both intra- and inter-person variability. The patient models were derived from data from real patients as detailed in the work by Patek *et al.* [82]. The sensor was considered to be noisy using the noise model proposed by Boyd *et al.* [17]. Each patient was run three times under three different noise scenarios. Nurses were assumed to commit no errors in typing in glucose values or adjusting infusion rates.

It was also assumed that checking blood glucose and adjusting insulin was done on the hour every hour, and that there was insignificant delay between when a blood glucose value was obtained and when the infusion rate was adjusted. The setup is shown in figure 4.15, where the decision-making software ($D$) is highlighted and the pieces that exhibit variability are shown.

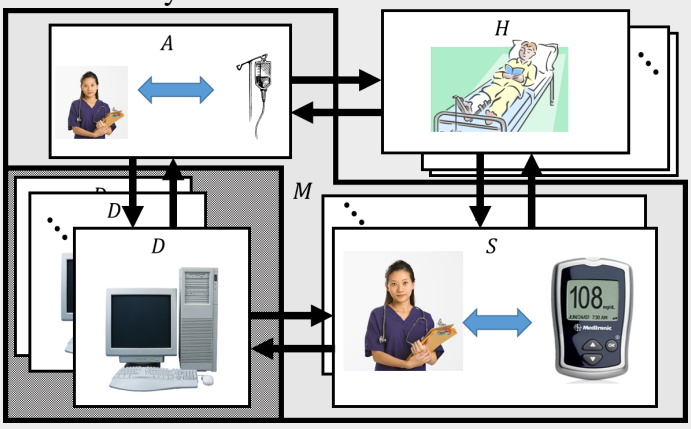



Figure 4.15: Scenario for assessment of insulin infusion decision-making software.

*Software Algorithm.* The software uses a basic algorithm whose intrinsic behavior is as follows. When the patient is first put on the protocol, no insulin is infused until the patient starts to exhibit hyperglycemia. Once infusion is started it follows this basic adaptive proportional control algorithm

$$y_{D\to A}[n] = K[n](\hat{x}_{bg}[n] - \underline{x}_{bg})$$
$$K[n] = \begin{cases} K[n-1] + \alpha & \hat{x}_{bg}[n] > \beta_{hi} \\ K[n-1] - \alpha & \hat{x}_{bg}[n] < \beta_{lo} \\ K[n-1] & \text{otherwise} \end{cases} \tag{4.28}$$

where $y_{D\to A}[n]$ is the insulin does in units/hr that the software instructs the nurse to give, $K[n]$ is proportional factor, $\hat{x}_{bg}[n]$ is the measured blood glucose value input into the software, $\underline{x}_{bg}$ is the blood glucose level at which infusion is stopped, $\alpha$ is the the amount by which the proportional factor is changed, $\beta_{hi}$ is a threshold glucose value above which the algorithm starts to increase the infusion rate, $\beta_{lo}$ is a threshold value below which the the algorithm starts to reduce the infusion rate.

For this particular case $\underline{x}_{bg}$ was set to 60 mg/dl, $\alpha$ was set to 0.01, and the initial value of the proportional factor ($K[0]$) was set to 0.02, based on values used by Steil *et al.* [98]. The thresholds were set to the target blood glucose thresholds ($\beta_{hi} = 180\text{mg/dl}$ and $\beta_{lo} = 80\text{mg/dl}$).

**Individual-Risk-Based Population-Level Risk.** The individual risks for each patient were computed according to equation 4.19 from the criteria in example 4.8. Since each patient had blood glucose trajectories ($x_{bg}(t)$) from three different trials, the maximum individual risk of the three trials for each metric was used as the risk for the patient. This means the functional risk and side-effect risk values for each patient could be based on different trials for that particular patient.

Once all 100 individual risks were available, we computed the population-level risk based on equation 4.20 from the criteria in example 4.8. The distribution data for the individual risks and the resulting population-level risk value for this version of the software (with the thresholds set to the glucose targets) is shown in figure 4.16.

**Acceptable Risk Assessment** According to the criteria from example 4.8, a population-level risk is acceptable if for each dimension (functional and side-effect), no more than 10% of the population of the population does worse than the baseline. Figure 4.17 shows the acceptable region (shaded area) for this case and the population risk

(the point (∗) shown) for the software from Figure 4.16. In this case the population risk is unacceptable, because it lies outside the acceptable region.

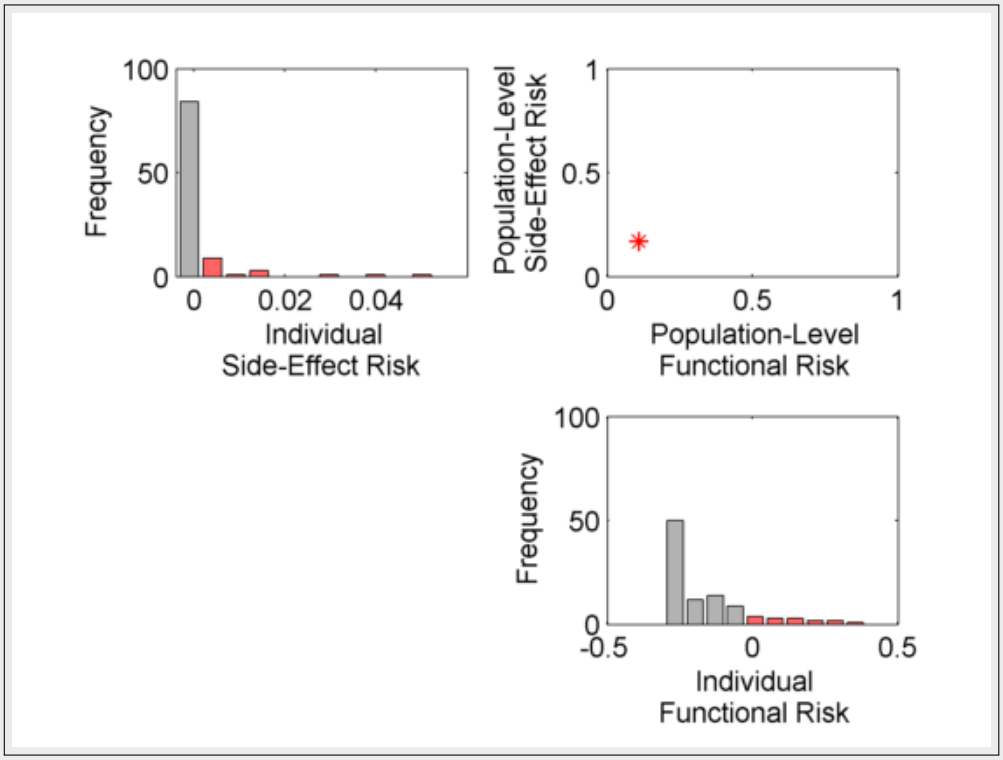



Figure 4.16: Visualization of the population-level risk and the distribution of individual risks.

Here, we are only concerned with the binary outcome. In a real design, this would prompt us to figure out why this particular outcome occurred, explore changes to the design of the software and reassess our changes. The next section on safety-guided design shows a case study where assessments are used to link risks to design features and develop requirements to ensure that system behaviors result in acceptable risks.

Note that the above is an assessment of the software, but based on the assumed (or expected) behaviors of the other parts of the health management system. Note also that this assessment is only valid for the way the health metrics were defined. Section 4.4.5 explores how changing definitions in the criteria like any of the metrics can impact the acceptability of the system under consideration.

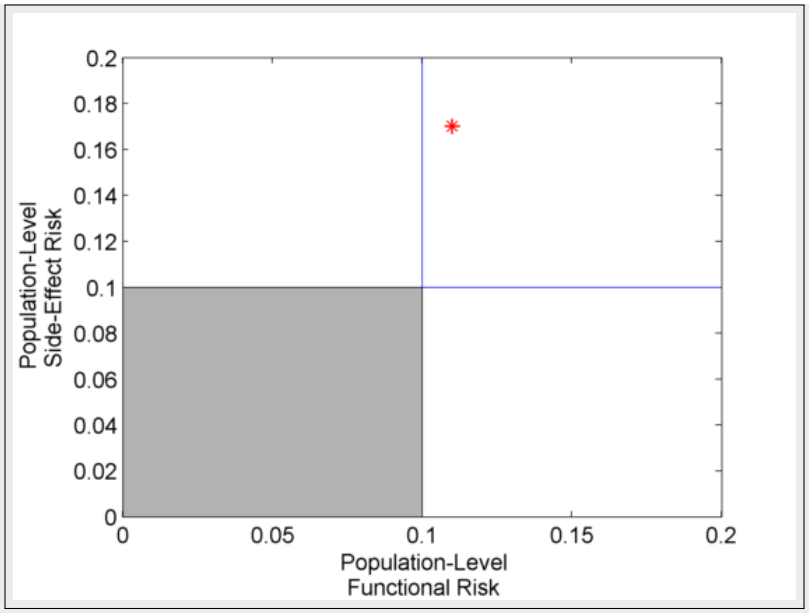



Figure 4.17: Individual-risk-based population-level risks for semi-automated insulin infusion protocol.

### Assessment of Multi-Use Designs

If a design is intended to be used in multiple distinct scenarios, the assessment could either be done for each scenario distinctly, or all scenarios can be considered jointly. In the joint case, the range of scenarios would represent the variability in the parts of the health management system not under consideration.

The choice of how to consider and assess the population-risk of the multiple scenarios really depends on the details of the health management scenario. One factor to consider is the type of risk metric used and the details of the baseline health metrics. If an individual-risk-based metric is used, then the joint case can be explored easily because each test health metric outcome will be compared to its corresponding appropriate baseline value. The direct risk metric can only be used if the baseline health metric values from the different populations are statistically similar. In the ICU case study for example, if it turns out that burn patients and pulmonary by-pass patients are stressed in similar ways statistically,

then health metrics from a collection of patients from both populations can be used for the baseline values.

The above presumes that the system we are considering does not change in any way across patients. For example, in the ICU case study, the assumption would be the the algorithm and configuration remains the same across all patients (or that any adaptation is not dependent on whether the patient is a burn patient or by-pass patient). In that case, considering the population risks jointly using a direct risk metric should work. Considering the case separately may reveal whether the system has lower risk for one patient population than another even if it is acceptable for both.

If the system does adapt in a population dependent way, then is must be assessed across patient populations separately and cases where the system is configured for the wrong patient type must be considered. In the ICU case study, for example, we would end up with some test health metrics for a scenario where system configured itself for a burn patients even though it was supposed to be interacting with a by-pass patient. Since the system is supposed to be interacting with by-pass patients, these test metrics would end up in the set used in population risk assessment for by-pass patients. A similar situation would be explored for the burn patient population.

If the adaptation is automatic, then the separate assessments would already implicitly account for the case where the system configures itself for the wrong type of patient. This is because when presented with a patient, the system must first figure out the patient type and configure itself according, and since we are doing a population-level risk assessment, the variability in the patients considered should test this capability of the system.

If the adaptation requires user input, however, then we would have to introduce misconfigurations intentionally. This could be done by introducing variability in the user behavior that allows this to happen. This would then factor into the population-level assessments just like the automatic adaptation case.

**Example 4.11:** Assessment of Artificial Pancreas for use in Adolescents and Adults

In this example we explore the case where the system is designed for different populations (adolescents and adults), but with no special adaptation for each population. The data for the test population comes from a simulation of an average adult and average adolescent. The baseline data come from that used by Kovatchev *et al.* which consisted of 55 patients who used a sensor and a pump but no decision-making software [52].

Inter-person variability was not explicitly explored in the test population. Rather the physiologic dynamics used for the patient are designed to represent the notion of an average Type-I diabetic based on the data used to create the glucose physiology model by Cobelli *et al.* [65]. The limitation to this physiology model is due to licensing reasons. However, this limitation makes the analysis simpler since the patient safety criteria developed for the artificial pancreas earlier ($mu_{M,1}$ given by equation 4.24) relies on the average health metric outcome for the population. The assumption therefore is that outcome based on the single model represents this average value for the test population. This is an example of using a proxy for a full assessment.

Figure 4.18 shows the population-level risk when both populations are considered jointly (marked by the star ($*$) symbol), for only the adult population (marked by the black dot ($\bullet$) symbol), and for only the adolescent population (marked by the green dot ($\bullet$) symbol). The acceptable risk region is defined by the shaded area (as specified in the criteria in example 4.9); the acceptable functional outcome risk is to the left of the vertical line and that for the side-effect risk is below the horizontal line.

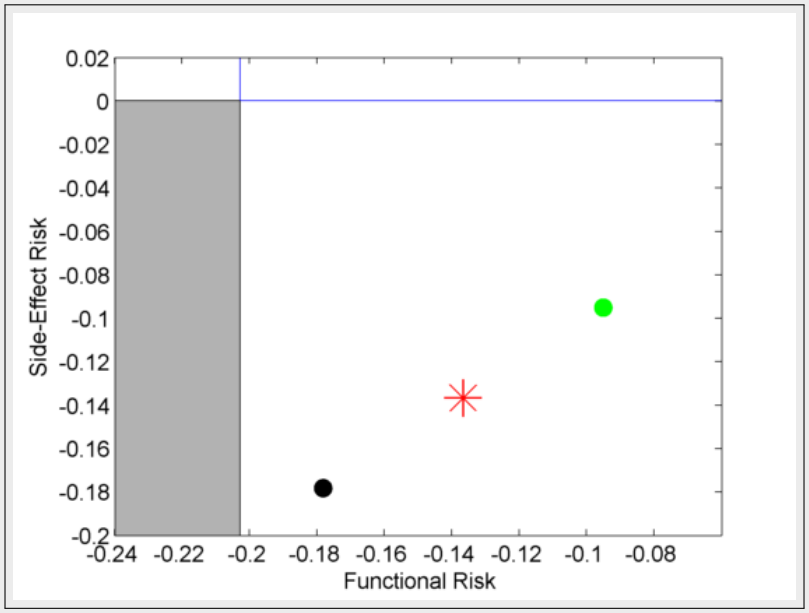



Figure 4.18: Population-level risk of artificial pancreas for adult (marked by the black dot ($\bullet$) symbol) and adolescent (marked by the green dot ($\bullet$) symbol) populations, and when both are considered as one population (marked by the star ($*$) symbol).

Even though overall the risk is unacceptable according to the criteria, the system seems to present lower risk for the adult population than the adolescent population.

### Comparison of Equivalent Technologies

Since the assessment of population-level risk only depends on the outcome of interaction of the technology with the patients, and not necessarily on the specific details of the technology, one can compute and compare the population-risks of (equivalent) technologies with similar functional goals (and hence functional metrics) and similar side-effects.

**Example 4.12:** Assessment of Equivalent Software for Semi-Automated Glycemia Management in the ICU

This example compares slightly different intrinsic behaviors of the decision-making software for the infusion protocol in the ICU. In previous examples, we assumed that the software sets the thresholds of reaction of the protocol the thresholds for the target blood glucose range. Here, we explore four other cases where the software sets the reaction thresholds ($\beta_{lo}$ and $\beta_{hi}$) some mg/dl away from the target blood glucose level thresholds. This results in the software reacting (and changing and proportional increment term ($K[n]$)) earlier than it would if the thresholds were the set to those in the previous examples.

Figure 4.19 shows the outcome of these comparisons. The range of behaviors were for threshold configurations $[+10,-0]$, $[+10,-10]$, $[+20,-10]$, and $[+20,-20]$, where $+\beta$ and $-\beta$ indicate the difference between the lower and higher reaction threshold and the target blood glucose thresholds respectively. The star ($*$) mark is original population risk from previous examples. The dot ($\bullet$) is the population risk based on the new intrinsic behaviors.

In general, it seems like reacting earlier at the higher threshold improves functional risk. It also seems like reacting earlier at the lower threshold improves side-effect risk, except when the software reacts much earlier for both thresholds.

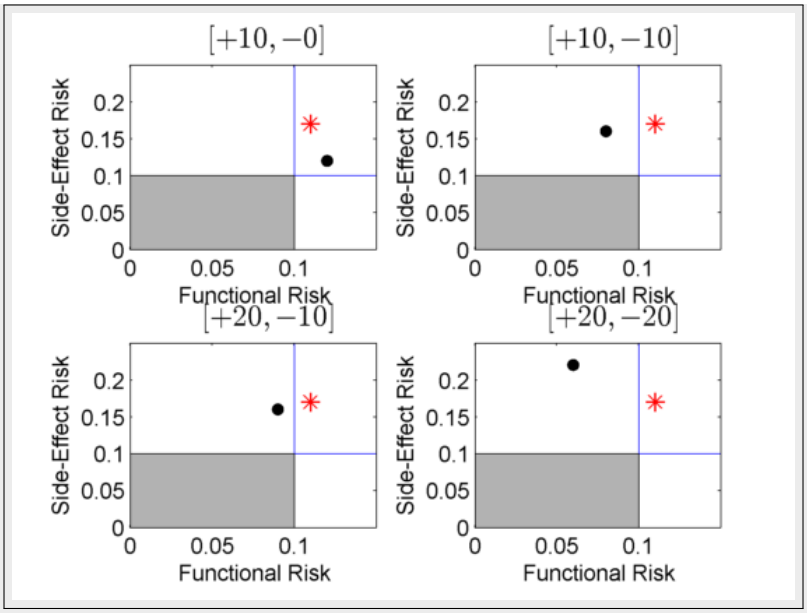



Figure 4.19: Comparison of four different intrinsic behaviors of the decision-making software for the insulin infusion protocol. $+\beta$ and $-\beta$ indicate the difference between the lower and higher reaction threshold and the target blood glucose thresholds respectively. The acceptable risk region is defined by the shaded area; the acceptable function risk is to the left of the vertical line and that for the side-effect risk is below the horizontal line.

In the example above, we explored slightly different intrinsic behaviors of the same system. This approach is particularly useful in a safety-guided design process discussed in the next subsection. We could have also explored systems with significantly different intrinsic behaviors but with the same required inputs and outputs. For example, we could compare two different pieces of decision making software that both take in blood glucose measurements and produce insulin infusion rate adjustments every hour but use very different algorithms for deciding on infusion rates. We could even have a case where one piece of software requires values more often (every 20 minutes, for example), or may instruct the nurse on when to provide the next blood glucose value update. A number of different algorithms can be found in the work by Steil *et al.* [98] where they evaluate the response properties (not patient safety implications) of a number of manual and automated insulin infusion protocols (including the one we have been using for our examples).

### A Note on Gathering Information Used in Assessments

Assessments require gathering a test set of metrics, and in some cases, also the baseline set of metrics. Assessments will be carried out at different stages of the design, and there are range of options on how to produce this data. At one end of the spectrum are minimal models, and at the other end are 'uncontrolled' trials as shown in Figure 4.20. The trade-off is between the cost (and feasibility) of obtaining the data, the realism exhibited by the data, and how 'rigid' the method of obtaining the data is.

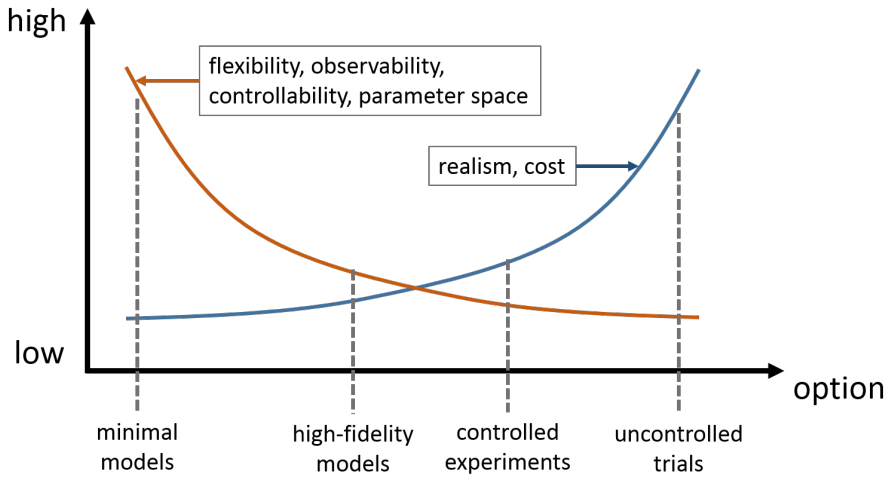



Figure 4.20: The spectrum of possibilities for generating health metric data for assessments.

Minimal (coarse) models are flexible and easy to use and manipulate, the also allow a wider range of explorations of the parameter space. They are typically used at the early stages of the design if available, though they can also be useful at later stages to step back from the details and answer questions about issues. 'Uncontrolled' trials are typically used in the final stages of the design, though they could also be used earlier if prototypes are cheap to develop and tests are cheap to run.

The inferences that can be made from minimal models are limited, and one cannot base a safety argument solely on them, though they useful (and sometimes the only feasible option) early in the process to help reduce the design space for when higher fidelity models or experiments are used to obtain data. In certain cases, especially exploring effects on health metrics of changing design parameters as is done in the next subsection, models

are the most feasible approach. The inferences that can be drawn from trials are stronger, though there is still the possibility that a smaller part of the parameter space is explored. Data from trials could also be used to improve models, or past data can be used to inform minimal model design for early concept analysis.

Recently there is interest in what are called *in silico* trials [81, 49, 46] where high fidelity models are used in place of controlled experimentation for explorations that would be infeasible in real experimentation, and also to reduce cost of the design process by eliminating experimentation like animal trials [49]. In some cases, the FDA is providing such models to aid manufacturers. An example is the virtual family model [**?**], which is set of anatomically-correct 3D models developed in collaboration with academic and industry partners for electromagnetic, thermal, acoustic, and computational fluid dynamics simulations, to examine safety issues like the use of MRIs on patients with implants. The FDA however makes no warranties about the reliability of the models and further states that using the model does not "[imply] endorsement by the FDA or [confer] any advantage in regulatory decisions" [109].

How to chose which process of collecting information is beyond the scope of this dissertation. What is important is understanding the trade-offs and strength of inferences that can be made and selecting the appropriate and feasible path. In presenting the safety argument (discussed later in section 4.4.6), one can provide rationale on choices of the particular option for collecting information.

### 4.4.3 Safety-Guided Design

In the design process, we would like the set of possible intrinsic behaviors of medical technology to result in acceptable population-level risks. One way to achieve this is to use what is called a safety-guided design approach. This is a top-down approach where the general principle is to first develop requirements and validate that any intrinsic behaviors that satisfy the requirements result in acceptable risks. The next step is to then design a system to meet the requirements, and verify and demonstrate that the design (and its implementation) meet the requirements.

In the end, the argument is that the resulting system results in acceptable risks because its intrinsic behaviors satisfy the requirements and we have already established that any system whose intrinsic behaviors satisfy the requirements must result in acceptable behaviors. The main idea behind this kind of approach is to have a traceable design process,

where each design feature is related to some set of requirements. Here, what we have done is add traceability between the safety requirements and the safety criteria.

A key part of this process is the establishment and validation of the requirements, and this is where the model is most useful. At the early stages, multiple potential designs can be assessed to understand the relationship between intrinsic behaviors of the system under consideration and population-level risks with respect to a particular (or multiple) patient safety criteria. With coarse models of (or information related to) the health management scenario under consideration, we can identify a set of black-box behaviors of the part of the health management system we are interested in that satisfy the patient safety criteria.

It is important to note that ideally, for each candidate behavior, a full assessment would be run. This could result in an explosion in the exploration space, which is a typical problem with design space explorations of this kind. One way to get around this is to use coarse but relevant models, as suggested earlier, for quicker exploration. Another way is to use some proxy for a full assessment, maybe based on knowledge from previous assessments. This would involve fewer assessments overall and estimating what the outcome of a full assessment would be from the fewer assessments. Lastly, since assessments are independent of each other, they need not be run in sequence. All assessments could in theory be run in parallel speeding up the time to obtain the information necessary.

Once candidate acceptable behaviors have been identified, resulting in narrower exploration space, higher resolution models (or more detailed information) could be used to ensure that the candidate behaviors do indeed meet the criteria. Once we have some level of confidence that the candidate behaviors are worth exploring, we can go about developing a more detailed design that produces these behaviors.

Figure 4.21 illustrates how the model can be used to guide the establishment of requirements which represent intrinsic behaviors that meet the patient safety criteria. The first part is developing a characterization of the intrinsic behavior of the part of the health management system we are designing. This creates an intrinsic behavior space ,$(\lambda_M^1, \lambda_M^2)$ in the figure, where $\lambda_M^1$ and $\lambda_M^2$ represent parameters related to properties of the system that can vary. For example, if we are designing a continuous glucose monitor, they could represent the point accuracy of a value reported by the monitor and the delay from when the value is sensed to when it is reported respectively. (Chapter 5 discusses how to determine the relevant intrinsic behavior parameters in more detail.)

The next step is to assess the population-level risk for multiple points in this behavior space with respect to an identified patient safety criteria. In this exploration, each point represents a test scenario where the intrinsic behavior of the system under consideration is

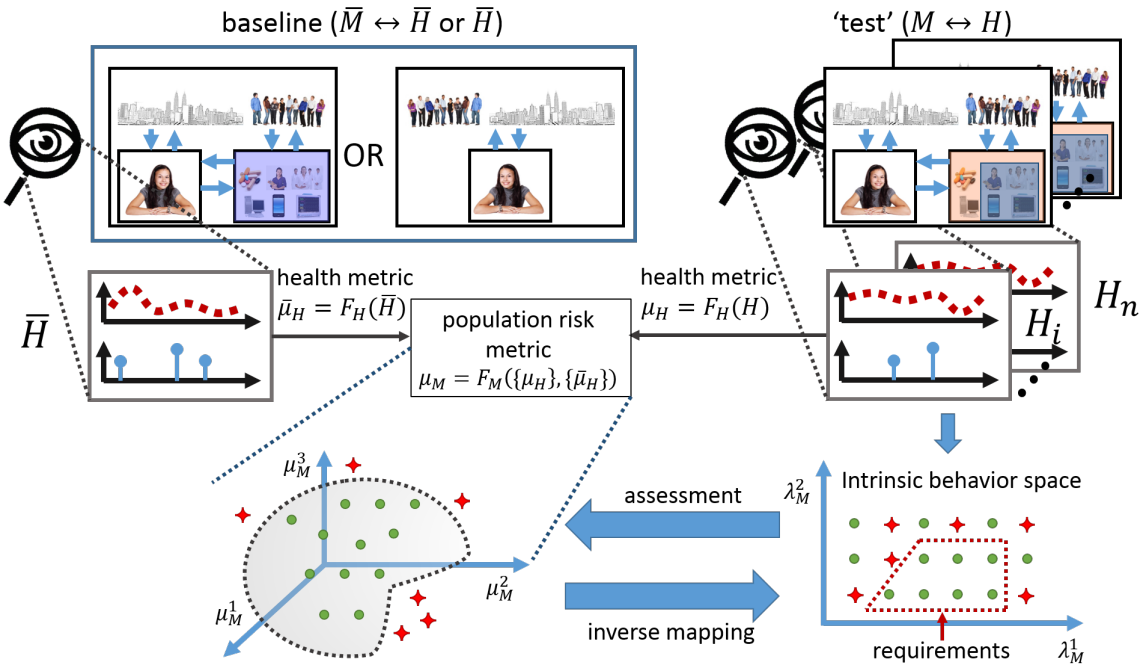



Figure 4.21: Conceptual depiction of establishing requirements for safety-guided design.

fixed but with variability in the other parts of the scenario. The assessment results in a set of points in the population-level risk space defined by the patient safety criteria subscribed to. Note that the risk space and the behavior space need not have the same dimensions.

Once the risk value of behavior points is established, the risk points that fall in the acceptable risk region can be identified. By doing so, this creates an inverse mapping of population risks to corresponding intrinsic behavior points, essentially tagging points in the behavior space as acceptable or not. Based on this information, an acceptable behavior region can be created in the intrinsic behavior space, and the criteria for this region becomes the patient safety requirements.

The approach described above essentially uses the acceptable risk region to create an acceptable behavior region. The description of this region, which would be a function of the parameters that make up the behavior space, becomes the requirements. Ideally, this region would be simple to describe and contiguous as in the figure (the region shown can be described by a trapezoid). Having a contiguous region helps provide some confidence that if the inevitable variability in the intrinsic behavior of the systems being designed is constrained to the acceptable region, the risk outcome will be acceptable.

Note that in the figure, because of the desire for simplicity in description of the acceptable behavior region, this region does not encompass all the acceptable points. If this region was mapped back to the risk space, it would result in a sub-region of the acceptable risk region. In general (and in this case) a direct mapping of the acceptable risk region to an acceptable behavior region could result in a behavior region that is quite complex to describe (and probably more difficult to design to).

The case study below looks at the establishment and validation of requirements for a meal sensing system for the artificial pancreas. The focus is on the artificial pancreas because there are more resources available to illustrate the utility of the model from this perspective. Focusing on meal sensing also demonstrate the use of the model in dealing with sensing subsystems (of which body sensor networks, which we will look at later, are emerging technology).

**Example 4.13:** Meal-Sensing-to-Decision-Making Interface Safety Requirements in the Artificial Pancreas

An important part of blood glucose management for Type-I diabetes is meal-time insulin infusion. The artificial pancreas requires meal information in order to infuse the proper amount of insulin at meal time. This information includes when the meal was taken and the amount of the meal that was taken.

Here, we are interested in designing a system that ensures that meals are reported in an appropriate manner to the decision-making software. In particular, we would like to understand what the interface requirement should be between this meal-reporting subsystem and the decision-making software. Since our aim is to gather requirements, the details of how this meal reporting subsystem works is not important. All we need to know from a black-box perspective is that the system interacts with the user (and possibly the glucose monitor) to produce meal information for the decision-making software to use.

**Characterization of the Subsystem.** The interface is based on what the decision-making software expects and what could happen in reality. This concept is illustrated in figure 4.22. In reality, the person consumes meals at different points in time which results in a meal trajectory ($u(t)$). The decision-making software finds out information

about these meals through what the meal-sensing system reports, which would be a meal information trajectory ($\hat{u}(t)$).

Details on the meal sensing system are omitted intentionally. It could range from a simple user interface requiring the user to manually enter information to an automated system that tries to infer meal information or something in-between where some inference is combined with user input. What we are most interested is the abstract version of this subsystem, which we can view as a subsystem that transforms the meal trajectory into a meal information trajectory as shown above the magnifying glass symbol.

In general, three things can happen to the meal timing information: it can be reported at time significantly before, right around, or significantly after the actual meal. For the meal amount information, the amount can be overestimated, underestimated, or just right. There is also the possibility that no information on the meal is reported at all. The requirements for the interface should specify what timing of report, estimation accuracy, and omission of meal information behavior is tolerable in order to meet the safety criteria.

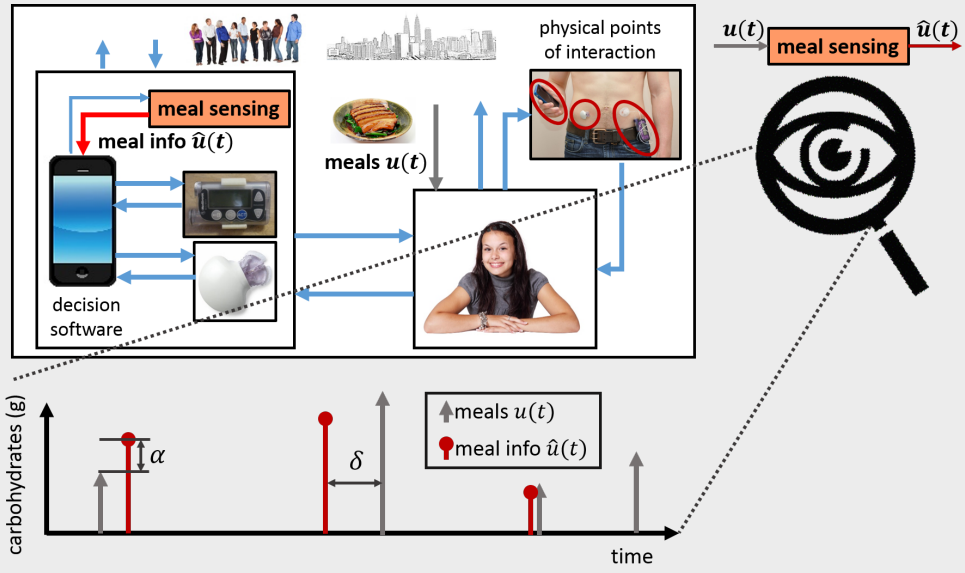



Figure 4.22: Illustration of meal sensing interface of the Artifical Pancreas showing information that the decision-making software expects and the possibilities between what is reported and what actually happens.

From the perspective of the interface, if we assume that a person eats three meals (breakfast, lunch, and supper), then each meal has an timing offset ($\delta_{i\in\{B,L,S\}}$, $B$ = breakfast, $L$ = lunch, $S$ = supper) which indicates how far in advance (negative offset) or how much later (positive offset) the meal is reported relative to when the meal is actually taken. Each meal also has a meal factor value ($\alpha_{i\in\{B,L,S\}}$) which indicates what the ratio of the estimated (reported) carbohydrate content in the meal to the actual value. Values greater than, equal to, or less than 1 indicate overestimation, perfect estimation, and an underestimation respectively.

Since, based on the above, each meal is characterized by a two-dimensional parameter space, an unreported meal can either be represented by an estimate of 0g of carbohydrate ($(\delta_i, \alpha_i) = (*, 0)$ where $*$ indicates that we do not care about that value) or a meal that is reported at an infinite time ($(\delta_i, \alpha_i) = (\infty, *)$). Accounting for all three meals collectively results in a 6-dimensional behavioral space for the interface ($[(\delta_B, \alpha_B), (\delta_L, \alpha_L), (\delta_S, \alpha_S)]$).

The space and a point in the space is visualized in figure 4.23. The space is visualized as collection of three two-dimensional spaces (one for each meal) a point in the overall 6-dimensional space is represented by the value of three points, one from each of the two-dimensional meal parameter spaces. In addition, a trajectory that corresponds to the point is also shown. Note that the behavior space definition does not state when a meal is actually taken.

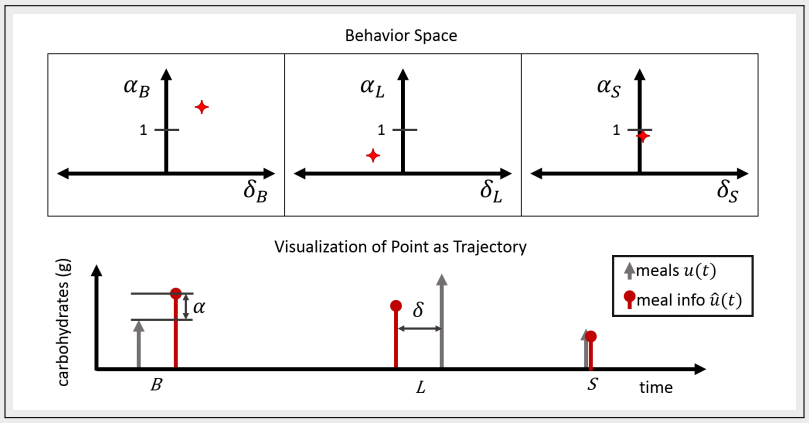



Figure 4.23: Illustration of the meal sensing interface behavior space and a point in the space. The point is combination of the points in each of the two-dimensional spaces shown.

**Assessment of Multiple Scenarios.** In order develop the requirements we must understand how variation in the intrinsic behavior of the meal-sensing-to-decision-making interface affects population-level risk. The requirements would limit this variation to only those that result in acceptable risk. To explore this variation, we must asses the risk of multiple scenarios, each representing one instance of the possible intrinsic behaviors of the interface. This was done through simulation of an average adult as done in example 4.11.

The health management setup is a same as that described in example 4.7. The glucose monitor was assumed to be noisy, following the model developed by Breton and Kovatchev [18], though actuations were assumed to be perfect. Also, if the patient did go into hypoglycemia, a standard a rescue dose of glucose would be provided (but the controller was not informed of this).

Each scenario consists of a day's worth of interaction (starting at midnight and ending at midnight). For each meal, five meal offsets ($\{-60, -20, 0, 20, 60\}$ all in minutes) and four possible meal factors ($4/7, 1, 1.5, 2\}$) and the case where no meal information was reported were explored. This results in 21 ($(4 \times 5) + 1$) points per meal and 9261 ($21^3$) scenarios. However, the case where no information was reported for all meals was omitted resulting in 9260 scenarios per person.

There was variability in the times at which meals were taken (hence the time between meals) and the amount of carbohydrate in the meal. This was introduced in the way meal offsets were actually implemented in the simulation. The meals we reported at the same time each day and the carbohydrate content of breakfast, lunch, and dinner were always reported as 65g, 95g, and 110g respectively. The actual meal was taken a certain offset (based on the offsets explored) from when the meal was reported, and the amount of carbohydrate was a certain factor (based on the reciprocals of the factors explored) of the reported amount.

The average outcome for the baseline, on the other hand, was computed from field data of a population of Type I diabetics who use a continuous glucose monitor and pump (in what is known as sensor-augmented pump therapy) but with no automated decision-making software. Since we had no inter-person variability data from the test scenarios, the overlap risk metric was not used.

**Identifying Acceptable Behavior Region and Developing Safety Requirements.** For the safety requirements, we are interested in statements about the parameters of

the interface. Particularly, we would like something of the form, meal offsets must lie between a certain set of values, meal factors must lie between a certain set of values, and the number of unreported meals must not exceed a certain number. This results in a region that is 6-dimensional 'box' with specific points left out.

To arrive at the requirements, we must first find the scenarios that produce risks that satisfy the safety criteria and find the common characteristics in terms of the parameters that allows us to find the edges of the box and the characteristics of the points in the box that must be left out.

The scenario exploration above essentially generates a list of records mapping a scenario (described uniquely by the values of the parameters) to a population-level risk for that particular scenario. We can therefore search this list to find which scenarios satisfied the criteria. Once we have those scenarios we look at them in terms of their parameter values to extract common characteristics.

To simplify the analysis, we take each meal parameter in turn, and consider the case of unreported meals separately. That is, we consider the meal offset across all three meals, and the meal factor across all three meals in turn, and not the individual dimensions of the 6-dimensional behavioral space. This corresponds to creating subspaces of the 6-dimensional behavioral space.

**Meal Offsets.** There are 3268 scenarios (out of the 9260) that satisfy the safety criteria. Figure 4.24 shows the distribution of the acceptable scenarios with respect to a number of different report timing classes. The first class (0) is when every meal is reported right around when the actual meal is taken. The second class (-20 to 0) is the case where every meal is either reported right around when the actual meal is taken or is reported no more than 20 minutes earlier than when the actual meal is taken (*i.e.,* a meal offset vector of the form $[\delta_B, \delta_L, \delta_S]$ where $-20 \leq \delta_i \leq 0$). The third class (0 to 60) is the case where every meal is either reported right around when the actual meal is taken or is reported no more than 20 minutes later than when the actual meal is taken (*i.e.,* a meal offset vector of the form $[\delta_B, \delta_L, \delta_S]$ where $0 \leq \delta_i \leq 20$).

The fourth and fifth classes (-60 to 0 and 0 to 60) are defined similar to the previous two. The last class (mixed) are scenarios that do not fall into any of the previous classes (*i.e.,* any of the three meals could be reported earlier, later, or right around when the actual meal is taken). The second to fifth classes represent consistency in

timing where if meals are not reported right around the actual meal, the are reported consistently earlier or later.

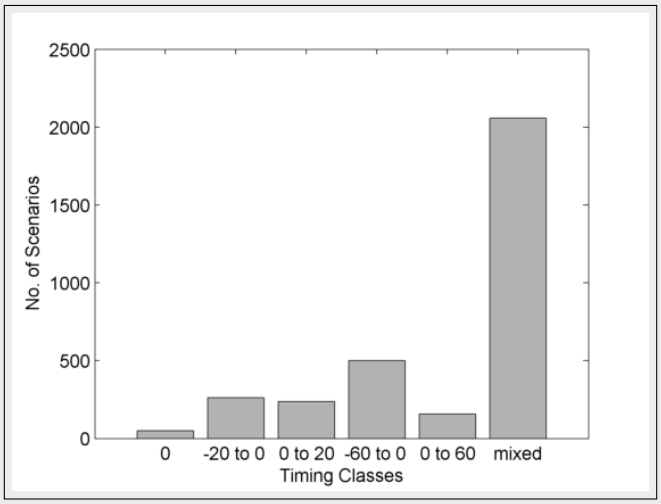



Figure 4.24: Distribution of acceptable scenarios with respect to meal report timing classes. '0' indicates that all meals were reported right around when the actual meal was taken, 'mixed' indicates that the meal offsets does not satisfy any of the other class definitions, '$\underline{\delta}$ to $\overline{\delta}$' indicates an offset vector of the form $[\delta_B, \delta_L, \delta_B]$ where $\underline{\delta} \leq \delta_i \leq \overline{\delta}$.

We would like our requirements to have some consistency so we can define edges of the box such that it contains only points that result in acceptable risk. Having the majority of the scenarios being in the mixed class does not help our cause. From the graph it looks like reporting meals more than 20 minutes after the meal was actually taken (0 to 60) results in fewer acceptable points.

Repartitioning the scenarios into those where no meal is reported more than 20 minutes after the meal was actually taken and those where at least one meal is reported more than 20 minutes after the meal was taken reveals the common characteristic of acceptable meals with respect to timing as shown in figure 4.25. Majority of the acceptable scenarios (close to 75%) are the case where no meal is reported more than 20 minutes after it was actually taken (-60 to 20).

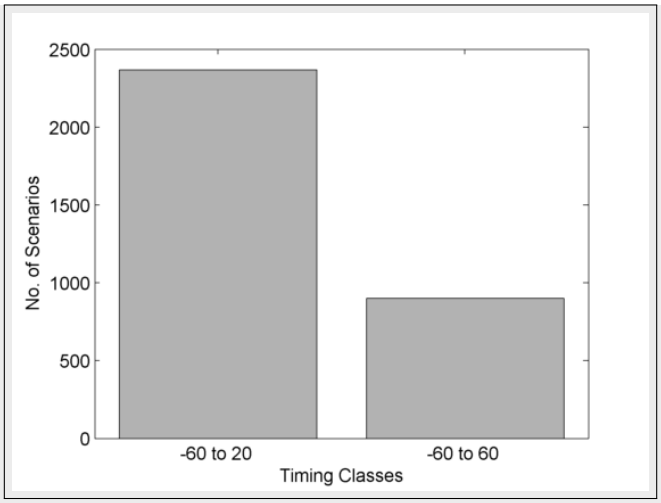



Figure 4.25: Distribution of acceptable scenarios with respect to meal report timing in the case (left bar) where no meal is reported more than 20 minutes after it was actually taken (but can be reported up to an hour before) and the case (right bar) where at least one meal is reported around an hour after it was actually taken.

**Meal Factors.** Figure 4.26 shows the distribution of the acceptable scenarios with respect to the classes of accuracy of carbohydrate content estimates. The first class (1x) is when all estimates are 100% accurate. The second class (1x or 1.5x) is when any meal estimate is either 100% or 1.5 times the actual carbohydrate content (*i.e.,* the meal factor vector is of the form $[\alpha_B, \alpha_L, \alpha_S]$ where $\alpha_i \in \{1, 1.5\}$).

The third and fourth classes are defined similar to the second class. The fifth class (mixed) is when a meal estimate can either be 100% accurate, 1.5 times, 2 times, or half the actual amount. It seems that overestimating the carbohydrate content in at least one meal while not underestimating the amount in any of the meals produces all the acceptable scenarios in the unmixed classes.

A deeper investigation shows that these scenarios (when the overestimation is 1.5 times or 2 times) account for the larger proportion (slightly greater than 70%) of all the acceptable scenarios as shown in figure 4.27.

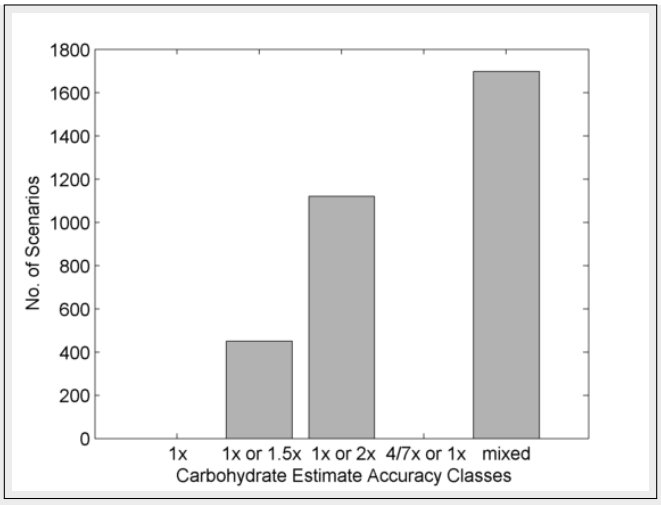



Figure 4.26: Distribution of acceptable scenarios with respect to classes of estimation accuracy of carbohydrate content in the meal. '1x' indicates that estimation was 100% accurate in all meals, 'mixed' indicates that the amount could have been overestimated or underestimated, '$\underline{\alpha}$ or $\overline{\alpha}$' indicates an meal factor vector of the form $[\alpha_B, \alpha_L, \alpha_S]$ where $\alpha_i \in \{\underline{\alpha}, \overline{\alpha}\}$.

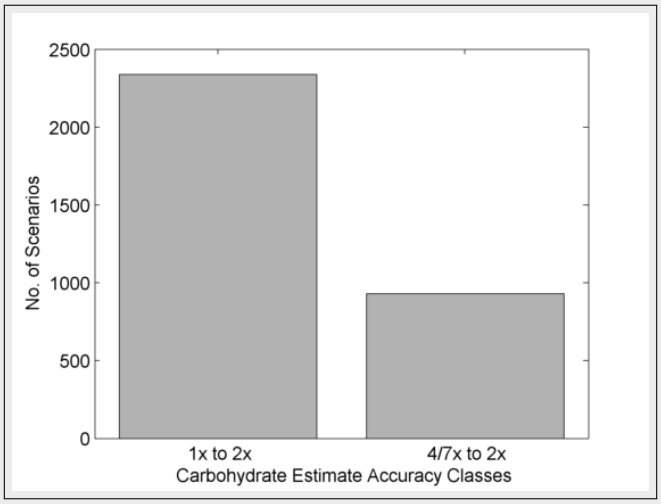



Figure 4.27: Distribution of acceptable scenarios with respect to classes of estimation accuracy of carbohydrate content in the meal. The left bar represents when the amount no meal is underestimated, and the right bar represents when the amount in at least one meal is underestimated.

**Unreported Meals.** Figure 4.28 shows the distribution of these acceptable scenarios with respect to unreported meals. Majority of the cases correspond to the scenario where no meal is unreported and none of the cases correspond to when two meals were unreported.

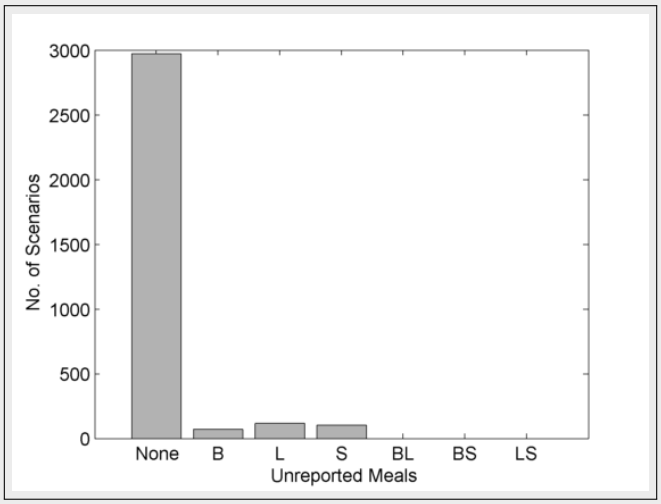



Figure 4.28: Distribution of acceptable scenarios with respect to unreported meals. B indicates breakfast, L indicates lunch, S indicates supper. The x axis indicates which meals were skipped and the y axis corresponds to the number of acceptable scenarios that correspond to those unreported meals

**Overall Requirements.** From the these results, we arrive at the requirement for the interface that a meal can be reported up to an hour earlier but no more than 20 minutes than it was actually taken (formally represented by Equation 4.29a), (oddly enough) no meal carbohydrate content should be underestimated but at least one meal estimate overestimated but by no more than 2x the actual amount (formally represented by equation 4.29b), and no more than one meal should go unreported (formally represented by Equation 4.29c).

$$-60 \leq \delta_{i\in\{B,L,S\}} \leq 20 \tag{4.29a}$$

$$1 \leq \alpha_{i\in\{B,L,S\}} \leq 2 \quad : \begin{bmatrix} \alpha_B \\ \alpha_L \\ \alpha_S \end{bmatrix} \neq \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{4.29b}$$

$$\text{Let} \begin{bmatrix} \omega_B \\ \omega_L \\ \omega_S \end{bmatrix} = f\left(\begin{bmatrix} (\delta_B, \alpha_B) \\ (\delta_L, \alpha_L) \\ (\delta_S, \alpha_S) \end{bmatrix}\right) : \omega_i = \begin{cases} 1 & \text{unreported meal} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Then} \sum_{i \in \{B,L,S\}} \omega_i \leq 1 \tag{4.29c}$$

**Validating the Safety Requirements.** The requirements form a 'box' in the behavior space, and hence is not a direct mapping of the acceptable risk region to the behavior space. As a result of this, we find (from our data) that some points that satisfy the requirements as stated are not acceptable according to the criteria we developed in example 4.9 (*i.e.,* there are points in the acceptable behavior region whose population risk values are not in the acceptable risk region). We also find that some points with acceptable risk do not meet our safety requirements (*i.e.,* there are population risk value points in the acceptable risk region whose corresponding behavior points are not in the acceptable behavior region).

Nevertheless, since requirements stated this way are easier to work with from a design perspective, we would like to know how bad the included unacceptable points are. Figure 4.29 shows the distribution in the risk space of the points in the acceptable behavior region (with the acceptable risk region highlighted).

First, only 20% of the points that satisfy the safety requirements are outside the acceptable risk region. Of these, none do worse than the baseline with respect to the average side effect outcomes (*i.e.,* if we look at only the side-effect dimension of the risk space, these points are in the acceptable region for that dimension). Hence all these points satisfy the side-effect risk criteria.

In addition, all the scenarios result in better average functional outcomes than the baseline and in majority of the cases (close to 85%), the results are at least 30% better (as shown in Figure 4.30). 'Relaxing' the safety criteria by accepting these points makes sense, especially since none of the scenarios are worse than the baselines on average.

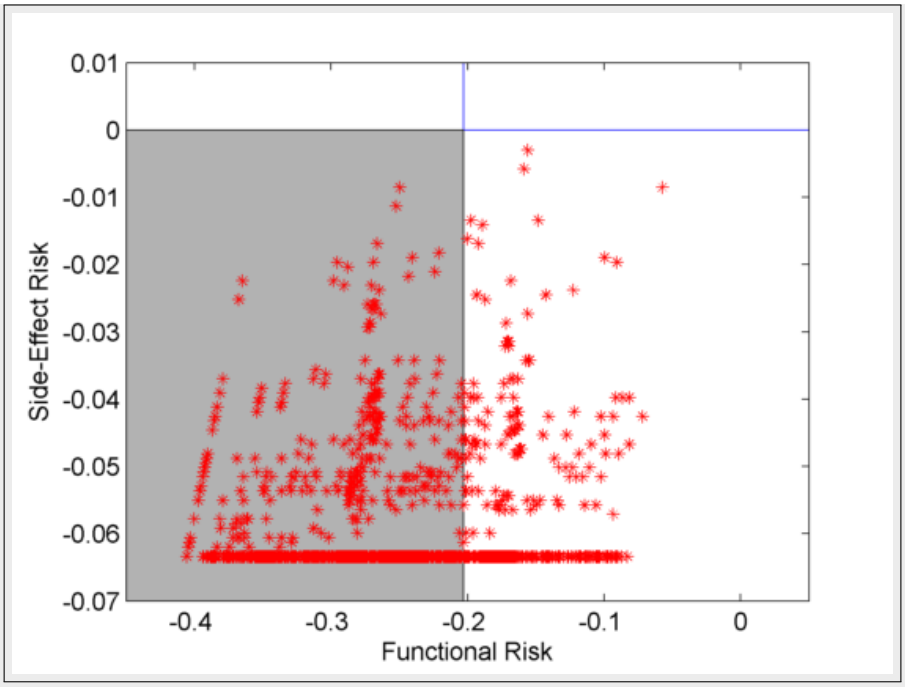



Figure 4.29: Distribution of acceptable behavior points in the risk space. The acceptable risk region is to the left of the vertical line.

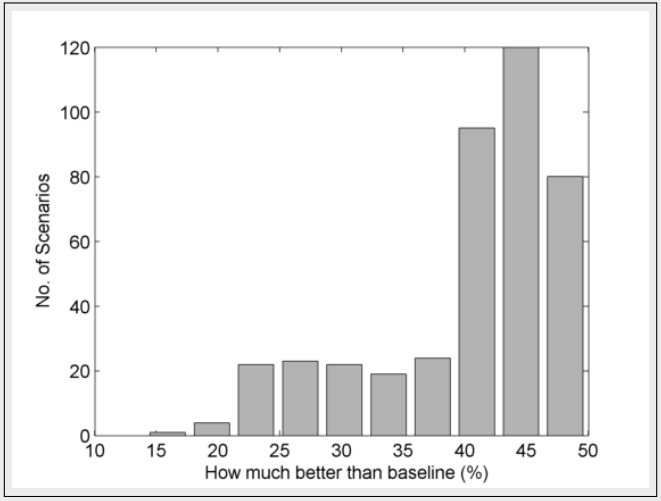



Figure 4.30: Distribution of scenarios the meet the safety requirement but are unacceptable according to the original safety criteria with respect to how much better than the baseline the average functional outcome is.

### 4.4.4 Discussion of Safety Criteria

When validating the requirements of the interface in the meal sensing interface case study, we relaxed the safety criteria so that points that were previously unacceptable from a risk perspective were now considered to be acceptable. This was done so we could end up with easier to state (and design to) safety requirements for the behavior of the interface. Our argument for the relaxation was that these points were close enough to the boundary of the acceptable region that accepting them made sense. This illustrates the utility of the model for discussing and adjusting safety criteria.

Safety criteria discussion can be motivated by many reasons, some of which include experience from the operational life of medical technologies, or new expectations from the patient population or health practitioners. Introduction of newer technologies to a an existing health management scenario could also prompt discussion of safety criteria. In some cases, for an emerging technology, the discussion may be prompted because a criteria does not actually exist. The manufacturer may have a discussion with the other stakeholders on what an appropriate criteria should be. One stakeholder group may propose a criteria and the others can agree to it or suggest modifications.

There may information available on the outcomes for the health management scenario to guide the discussion (like in example 4.13) or all that may be available would be past data for a similar scenario. Nevertheless the stakeholders, using the model, can be explicit about the assumptions and rationale that goes into the criteria that is agreed upon. The next section discusses how the model can be used to explore the impact of different assumptions and rationale (choices of components of the safety criteria) can affect what is considered acceptable both in terms of risk and in terms of behavior.

### 4.4.5 Exploration of Impact of Assumptions and Rationale

One of the major benefits of the model is the ability to explore the impacts of assumptions and rationale on the patient safety criteria and the hence the design requirements. Since the patient safety criteria is linked to all the choices made along the way (the assumptions about different parts of the system—including the patient population, the health metrics, the risk metrics, and baselines), we can explore the effects of changes of the choice of any of these on the overall assessment outcome (and requirements determination).

Here, we are interested in the case where for a *given* system design, we would like to know how assumptions made about the different parts of the safety criteria affect what

conclusion we draw about its patient safety, and not the assumptions about the behavior of the system itself.

This exploration can be used as part of a sensitivity analysis to understand what might happen if some of our assumptions (about populations, health metrics, and other parts of the health management system not under consideration) were actually wrong. It could also be used to compare results of assessments based on older assumptions to assessments informed by new data available resulting in newer assumptions. We could also conceivably use information from data gathered in the field to select a set of assumptions that best fit the data.

Below the impact of assumptions on risk assessment outcomes are illustrated looking at changes in each of the pieces of the safety criteria in turn.

### Health Metrics

The example below demonstrates the impact of the choice of health metrics on the overall risk assessments. From the perspective of the model, the choice of health metrics ($\mu_H$) affects the definition of the axes of risk metric ($\mu_M$) since it can affect the possible range of values of the health metrics and hence the risk metric. For the same risk metric, changing the health metric may not change the interpretation of the outcome, but may change the conditions on which the interpretation holds as shown in the example below. In general, changing the health metric will require changes to some aspect of the intrinsic behavior of the medical technology, since the medical technology is designed to help improve the outcome of at least the functional health metric.

**Example 4.14:** Impact of Health Metric Choice on Risk of Semi-Automated Glycemia Management in the ICU

This example explores different target blood glucose ranges for the insulin infusion protocol, and hence different definitions of normal for the functional risk metric. The target blood glucose ranges were taken from the Portland Protocol (four different versions) [84], and the University of Washington Medical Center protocol and the Yale-New Haven Hospital protocols as described in the work by Steil *et al.* [98]. We used the original 100 patient population (where all physiologies are equally likely) and the software version where the thresholds are set to the target blood glucose values.

Figure 4.31 shows the population level risk for each of the protocol versions. The star ($*$) mark is original population risk from previous examples. The dot ($\bullet$) is the population risk based on the new health metrics. Even though the normal ranges for the health metrics (original and new) are different, both population risks can be plotted on the same axis because the population risk metric is defined as the percentage of the population that does worse than the baseline, and does not depend directly on definition of the health metric: regardless of the definition of the health metric, the risk metric as defined will always range from 0 to 100% (0 to 1 on the normalized scale used in the figure).

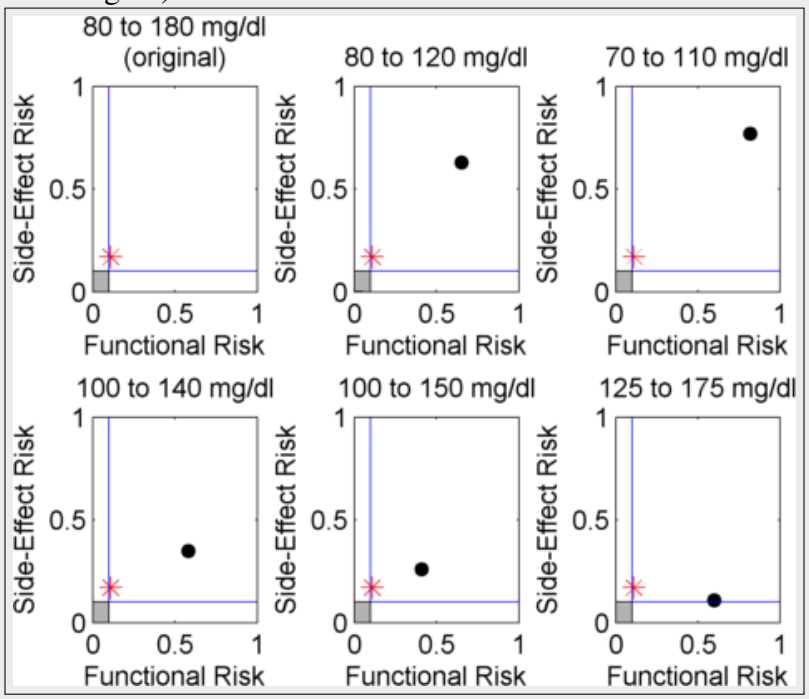



Figure 4.31: Population-level risk for semi-automated glycemia management in the ICU for two different assumed populations

If it turns out that our original target range is not consistent with other longer term outcomes used to define these ranges, then the actual risk of our system could be different (higher in this case) as seen in the plots in the figure. It seems like the intrinsic behavior of the algorithm is best suited to the original target range of 80 to 180 mg/dl, and it would have to be adjusted to accommodate newer ranges.

### Health Management Scenario

The example below demonstrates the impact of the assumptions about the population the overall risk assessments. From the model perspective, assumptions about the population will affect the set of health metrics (both baseline ($\{\overline{\mu}_H\}$) and test($\{\mu_H\}$)) which are passed to the population risk metric.

**Example 4.15:** Impact of Population Assumptions on Risk of Semi-Automated Glycemia Management in the ICU

This example explores the case where the nature of the patient physiologies assumed may be different from what is appropriate. In the previous examples, since each patient physiology was represented only once, the assumption was that each physiology was equally likely. Here we explore the case where some physiologies are more likely than others.

We generated another population from the original set where some physiologies were more likely than others as follows. Each patient's data is associated with a patient ID, and hence the physiology is represented by the ID. Since there were 100 patients, there were 100 IDs. We generated the two sets of 200 patient IDs (1 to 50 and 51 to 100) each based on a normal distribution, creating a set of 400 patient IDs where values ranged from 1 to 100 and some numbers were repeated more often than others. The resulting distribution of patient physiologies (represented by the patient ID number) is shown in figure 4.32.

We then computed the population risk for the new 400 patient population following the same procedure as before in example 4.10. The resulting population-level risk is shown in figure 4.33. The star ($*$) mark is original population risk from previous examples (100 patient physiologies all equally). The dot ($\bullet$) is the population risk from the newly generated population (400 patients based on 100 physiologies with some physiologies more likely than others).

If it turns out that our original test population was not an appropriate representative population, and this newer population is more representative, then the system would have a higher functional outcome and side-effect risk than we originally thought.

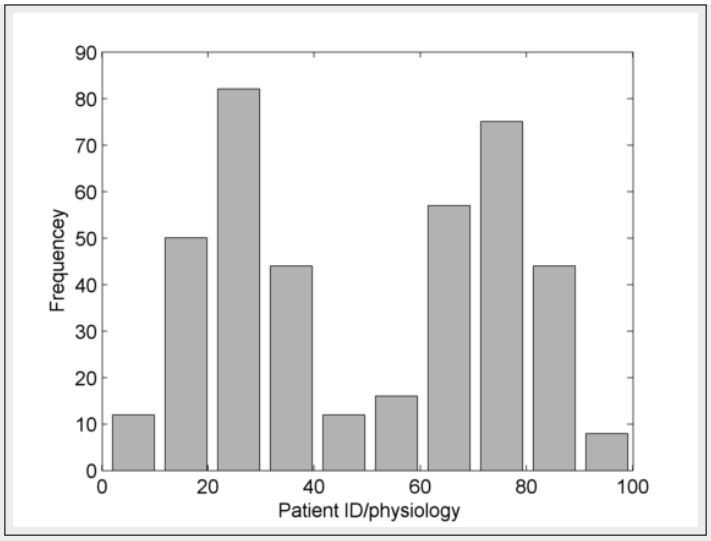



Figure 4.32: Distribution of patient physiologies (represented by the patient ID number) in new assumed population

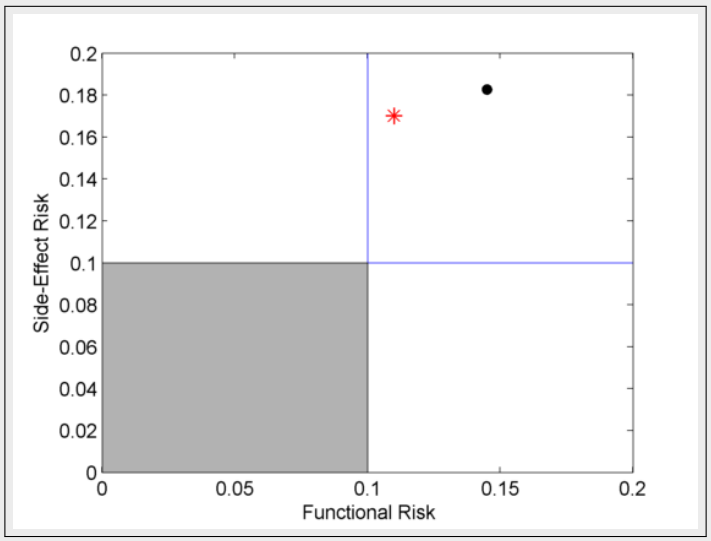



Figure 4.33: Population-level risk for semi-automated glycemia management in the ICU for two different assumed populations. The star (∗) mark is original population risk from previous examples (100 patient physiologies all equally). The dot (•) is the population risk from the newly generated population (400 patients based on 100 physiologies with some physiologies more likely than others). The shaded area is the acceptable risk region.

### Population-Level Risk Metrics

The example below demonstrates the impact of the choice of population-level metrics on the overall risk assessments. From the perspective of the model, the choice of population-level risk metrics ($\mu_M$) affects the definition of the axes of risk metric and the interpretation of the results. In particular, the range of possible values of the outcome of the risk metric may be different for different definitions.

**Example 4.16:** Impact of Population-Level Risk Metric Choice on Risk of Semi-Automated Glycemia Management in the ICU

This example explores a different population-level risk metric than the original. The new metric still uses the same individual risk metric defined in equations 4.12 and 4.13, but defines a 6-dimensional population-level risk metric based on the mean, minimum, and maximum value from the set of functional and side-effect individual risk values. This can be visualized as a dot (for the mean) and a box using the minimum and maximum values for the length of edges in each direction as shown in figure 4.34.

The acceptable risk region is defined in this case by a mean equal to the baseline (a value of 0), and minimum value of of the minimum possible individual risk metric value (-0.3 in the functional outcome case, and 0 in the side-effect risk case), and a maximum of 0.1 for the functional outcome risk case and 0.01 for the side-effect risk.

We computed this new population-level risk for the original 100 patients (all physiologies equally likely) and the original behavior of the decision-making software. The test set of health metrics and individual risk metrics were the same as computed in example 4.10. The only difference was the population-level risk.

Figure 4.34 shows the population level risk for the original metric definition (on the left) mark and the new metric (on the right). Notice that the axes are different in the range of values, even though all the data input to both metrics (the set of individual risk values) are the same. In particular, the functional population risk in the new metric can have negative values, whereas in the original metric both risk values could only lie between 0 and 1 inclusive. Also, whereas the original risk value consists of a point in two-dimensional space, the risk value for the new metric consists of the point and the (white) box surrounding it.

One way to compare the choices is to look at the acceptable risk regions and to determine if the population risk for this same system is acceptable under one metric and not acceptable in another. In both cases, the risk is unacceptable because in both cases the 'points' lie outside the acceptable risk region. In the new risk metric case, even though the mean value is within the region, the range of values (determined by the box with black lines) lie outside the region.

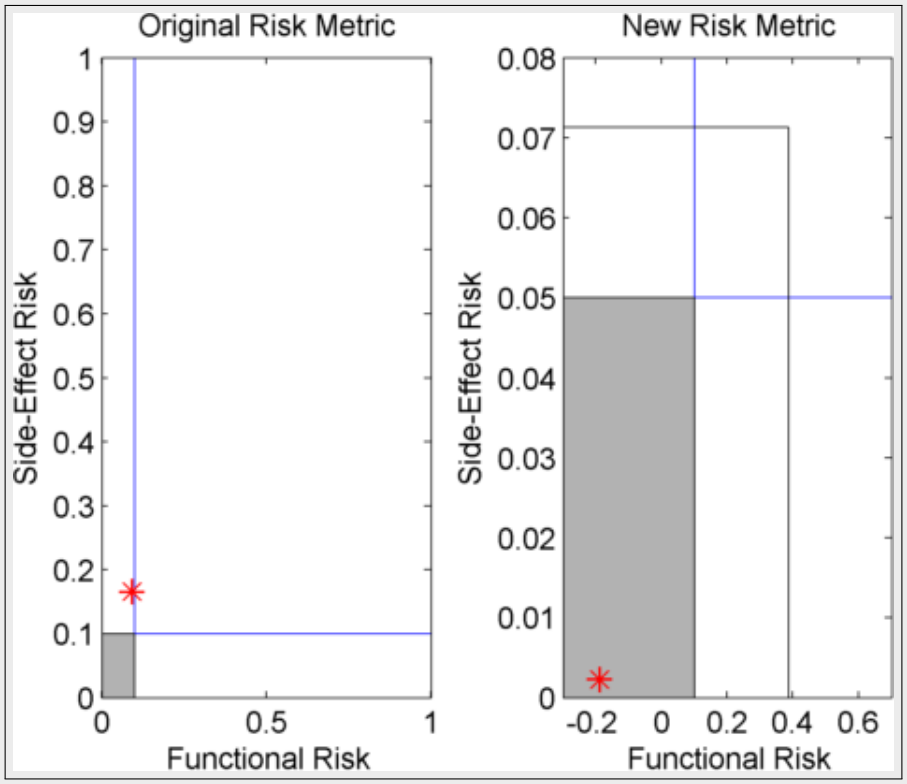



Figure 4.34: Population-level risk for semi-automated glycemia management in the ICU for two different assumed populations

### Acceptable Risk Criteria

The example below demonstrates the impact of the choice of acceptable risk criteria on the overall risk assessments. From the perspective of the model, the choice of the acceptable risk criteria ($\overline{\mu}_M$) changes the acceptable risk region and hence the health management

scenarios (or behaviors of the subsystem under consideration) which are considered acceptable.

**Example 4.17:** Impact of Acceptable Risk Criteria Choice on Risk of the Meal Sensing Subsystem for the Artificial Pancreas

This example shows the impact of the relaxation of the safety criteria that was done in example 4.13. Figure 4.35 shows the population level risk for the original criteria definition (on the left) and the new criteria (on the right). The star ($*$) marks indicate the unacceptable risk values and the dot ($\bullet$) marks indicate the acceptable values. Notice that the acceptable risk region is larger under the new criteria than under the old. Also, the new acceptable risk criteria is irregular (as opposed to rectangular in the original criteria), because it is based on the acceptable behavior region (and not independent of it as would usually be the case).

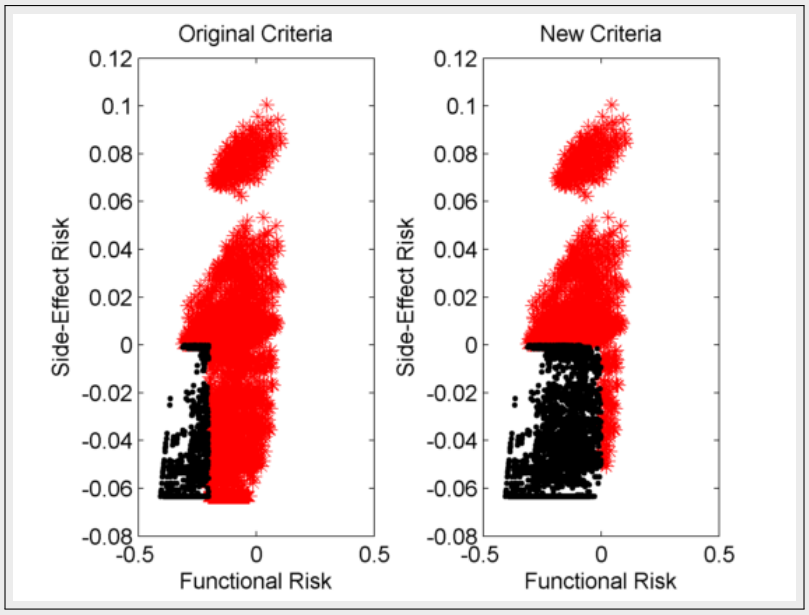



Figure 4.35: Population-level risk for artificial pancreas for two different acceptable criteria

If two different stakeholders have different conceptions on what the acceptable criteria is, explicitly defining it using the mechanisms provided here allow them to have a discussion and come to a common conclusion. However, if this is left implicit,

it is more difficult for them understand why one might consider a certain system acceptable when another considers it unacceptable.

### 4.4.6 Safety Argument Structure

In chapter 2, we brought up the issue of regulation for safety-critical systems where the regulator must make a decision on whether the system presented by the design must be allowed to operate. The process involves the the designer presenting information on the design, the design process, and the results of the assessment, and making an argument that the information supports the inference that the system would behave safely in operation.

The structure of the safety criteria provides a way to structuring safety arguments. By making all assumptions in assessments explicit, it allows the regulator to question all assumptions and rationale. In effect the designer is arguing first that the patient safety criteria used is valid, that the data used in the assessment (which includes the range of issues explored) and the process that was used to gather it is valid, and if we accept the validity of these two things, and the assessment shows their system to have acceptable risk, then it should be approved as safe. At this point, the designer could also argue that the criteria they were given (by the FDA) is too stringent, and needs to be relaxed. Conversations can be had with the FDA and health practitioners on what a new criteria might be.

In addition, the designer need not wait till the end of the design process before making an argument. Early in the process, they can (and should) have conversations with the other stakeholders to establish a base criteria before beginning the design process. Throughout the process, conversations could be continued to refine the criteria along with the design if new information suggests this, so that all stakeholders are on the same page about the criteria used to inform design throughout the process. This can save costs since changes later in the design process are more costly than earlier.

## 4.5 Relationship to Other Medical Technology Safety Work

As mentioned previously there are some efforts in safety assurance for medical technologies. The generic infusion pump project by the FDA in collaboration with a number of academic partners [47], for example is focused on providing a reference model for infusion pumps that manufacturers can as a starting point for designs. It provides a set of hazards and a formal model of the infusion pump software that the group continues to refine and

verify that is hazard-free. This work makes no links to expected patient outcomes. It assumes that the requirements part of the safety guided design process described in section 4.4.3 has already been undertaken and the acceptable behaviors have been identified. The formal model, in effect is a model of a system that exhibits the acceptable behaviors.

Another approach is work by Jiang *et al.* on pacemakers and Pajic *et al.* on a networked closed-loop system [46, 78]. Both use patient models to provide analytic guarantees that certain behavior of the closed-loop algorithms will ensure patient safety, defined as vital signs within some range. Both acknowledge variability as an issue to consider, but do not seem to incorporate this explicitly in their approaches. Neither really focus on how the patient safety criteria as it pertains to patient outcomes is derived or expressed. They take the hazards or vital sign regions as given and mostly explore worst case scenarios of system behavior, implicitly defining a health metric (the vital signs) and an acceptable risk region (vital signs should never go out of a particular range).

These implicit approaches are a poor way to account for variability, and also represent a more system-behavior-space-oriented approach as opposed to the patient-outcome-centric approach advocated in this dissertation. Though this is approach is well suited to design, it leaves many assumptions about the patient safety criteria implicit. The work presented here can complement these approaches by helping to determine a more explicit patient safety criteria as well as requirements (an acceptable behavior region) for safety-guided design. Then the above approaches can be employed to ensure that proposed designs do indeed fall within the acceptable behavior range.

The closest work to what is presented here is that for a specific technology by Kovatchev *et al.* [49], Magni *et al.* [64] and Patek *et al.* [81] for the artificial pancreas. In particular *in-silico* trials on a variety of patients (developed from real data) are used in order to explore the variability of patient outcomes. They use what is effectively an individual-based population risk approach called the control variability grid analysis (CVGA) [64]. The health metric is a two dimensional metric consisting of the minimum and maximum value of the blood glucose trajectory (in mg/dl) of the patient while using the artificial pancreas for a day. Eating behavior of the patient is also simulated.

Though they do not define a baselines the CVGA defines a population risk 'metric' by considering 9 regions where the health metric values can fall. Four of these regions are considered the 'safe' regions and five of them are considered the 'control error' regions. The risk metric defined is actually a three dimensional metric consisting of the proportion of the patient population who have health metric in the 'safest' region, 'safe' regions, and 'control error' regions.

In addition, in recent work on safety requirements for continuous glucose monitors [52], Kovatchev *et al.* use this *in-silico* trials with variability approach to essentially identify acceptable behaviors similar to what was done in the section on safety-guided design.

The artificial pancreas work does not explicitly define the population risk metric or population risk space, though this can be done easily. It also does not develop an acceptable risk criteria, it is implied that the risk is to be low as possible, but not clear how low (or how many errors are acceptable). Both works (CVGA and that on glucose monitors) are a special cases of the general patient safety model presented here with a few pieces missing. The patient safety model in this chapter makes the notions more explicit.

## Summary

Medical technologies are an essential part of health management, whose goal is to reduce the natural (health) risk associated with human function when this risk becomes unacceptable. However, because of the variability associated with the patients who these technologies must help (and other parts of the health management system), they can fail to achieve their intended goal or introduce new risks when they interact with patients. It is important to express and evaluate this technology risk in order to decide on what is acceptable and inform system design.

This chapter provided a patient-outcome-centric way of expressing this risk, based on the notion of health metrics as a human function risk measure developed in chapter 3. It showed that this makes safety discussion for medical technologies more focused on their intended goals, and demonstrated how this model of patient safety can be used for various aspects reasoning about the patient safety of medical technologies. The model presented here generalizes recent efforts for emerging technologies like the artificial pancreas and complements other efforts focused on verifying (and guaranteeing) the designs exhibit the intended acceptable behaviors.

# Part II

# Implications and Applications

# Chapter 5

# Implications for Safety Analysis of Body Sensor Networks

> According to Goloumb, "you will never strike oil by drilling through the map!" But this does not in any way dimish the value of the map.
>
> Edward A. Lee

## Chapter Overview

Chapters 3 and 4 discussed some of the subtleties associated with reasoning about sensing subsystems because of some of their indirect effects on the human dynamical system. Chapter 4 showed how requirements could be developed for sensing subsystems by defining a interface between sensing and decision making. This chapter looks at this further, showing what properties must be considered as part of the interface and also considering issues in interactions with the patient. It focuses particularly on body sensor networks (BSNs), an emerging class of medical technologies for sensing for health management that embody the three complexity-increasing trends of integration, autonomy, and mobility.*

*The main ideas in this chapter (the generic body sensor network model, the generic set of hazards, and the causal factors) were original developed in a paper presented at the International Conference on Body Area Networks [9]. The ideas have since been updated and what is is presented here is the updated version.

## 5.1 Introduction

Body sensor networks (BSNs) present a unique opportunity for improving the quality and mobility of healthcare. Such systems enable patients to continue their normal daily lives and 'invisibly' collect patient information under dynamically changing environments. This enables healthcare practitioners to access otherwise-unobtainable information to assist and improve medical decision-making, and to gain better understanding of how the human body functions in various environments. Realization of the BSN vision will significantly influence both medical research and practice, as evidenced by a number of preliminary studies highlighted in the survey article by Pantelopoulos and Bourbakis [79].

The ultimate challenge is to assure the safety of patients who use BSNs. Even though BSNs do not directly deliver treatment or medication to patients, they do collect and supply information that, when used in medical decision-making processes, have significant impact on the correctness of decisions made, and hence on the patient's safety. Assuring patient safety is especially challenging in BSNs because of their differences from their in-clinic counterparts. BSNs are typically governed by more stringent constraints on their resource consumption (e.g., energy and computational resources), as well as mobility and device size constraints (see reference [21] for discussion of such issues). More importantly, BSNs are operated in scenarios typically outside a clinical environment. Therefore, access to medical practitioners and service technicians is usually limited.

This chapter builds on the foundation provided in chapters 3 and 4 to develop mechanisms for reasoning about the patient safety of BSNs. First, a generic model for BSNs is developed to define the system scope for BSNs. This model captures the conceptual role of the BSN in health management as well as the nature of its interactions with other entities in the health management scenario.

This model helps us identify hazards at the interfaces between the BSN, the human and their environment, and the rest of the health management system. The behavioral properties of BSNs that factor into these hazards form the behavior space that must be explored as discussed in section 4.4.3. Patient safety for the BSN is defined in terms of these interface hazards (since safety requirements can be defined in terms of an acceptable behavior region).

Lastly, in the spirit of systems safety engineering, the generic model is used to discuss causal factors for the hazards and a (non-exhaustive) list of points to consider in order to inform safety-guided design. Examples of how ideas map to real sensing scenarios are provided to show the validity of the ideas.

## 5.2 The Generic Body Sensor Network Model

The main aim of a sensing subsystem like a body sensor network in the health management system is to provide information of interest for decision-making. This apparently simple process is wrought with a number of complexities. Figure 5.1 shows a conceptual picture of the interactions involved in achieving this aim. The actuation interactions with the patient are deemphasized since these are not the primary concern of the BSN. Though they are deemphasized, keeping them in mind is important since the behavior of the BSN does indirectly factor into those interactions.

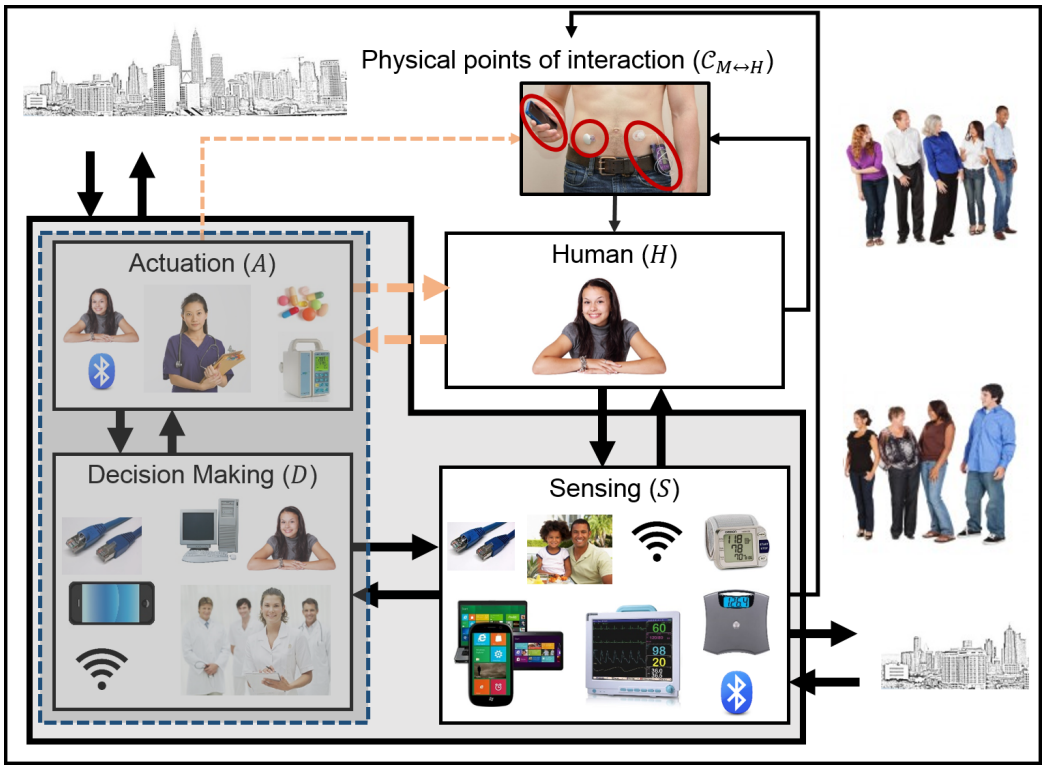



Figure 5.1: Conceptual view of the generic body sensor network model

In the rest of the chapter, the more abstract view (shown in figure 5.2) of the conceptual scenario shown in figure 5.1 will be used for developing the various ideas. The component $H$ represents the patient and their environment. In previous representations, all components are embedded in a larger environment, but since most of the BSN is usually physically coupled to the patient, it is more convenient to consider the patient's environment part of

$H$. The component $S$ represents the BSN, and the composite component $D,A$ represents the decision-making and actuation components. From the perspective of the BSN, the details of that component are not important. Note, however that since this is a conceptual view, if the patient is the decision-maker, then they are represented both by $H$ and $D,A$.

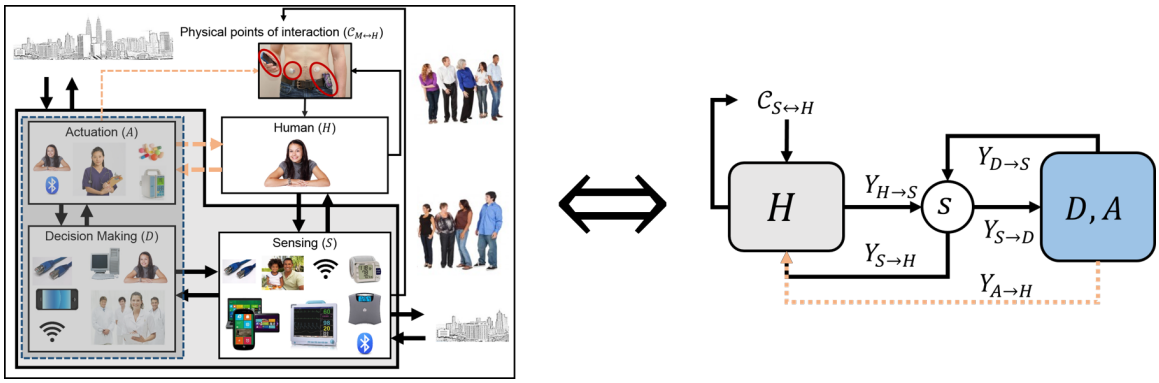



Figure 5.2: Abstract view of the generic body sensor network model

The arrows indicate the interfaces between various components. Various physical quantities or abstract messages (the $Y_{(*)\to(\cdot)}$ values) can be exchanged on these interfaces and the arrows indicate the direction of flow of these quantities. Components may produce an output on an interface 'spontaneously' or in reaction to an input received on another interface. Below the nature of the interfaces and the quantities exchanged on them are described and discussed in order to provide the mechanisms for identifying hazards and discussing causal factors. The actuator interface to the patient is omitted because that interface is not part of the BSN and out of design control of the BSN designer.

### 5.2.1 The Coupling Interface ($\mathscr{C}_{S\leftrightarrow H}$)

Part of the BSN must be coupled to the patient (which in this case includes the patient's environment) so that the BSN can interact with the patient to gather the information needed by the rest of the system. The nature of this coupling is captured in the coupling interface ($\mathscr{C}_{S\leftrightarrow H}$). This could include information on the physical location of the BSN components on the body (*i.e.,* the points of contact with the body). For example, in electrocardiography, electrodes are usually placed on the patient's limbs and chest. It could also include information on the nature of the physical contact. For example, how loosely or tightly are the electrodes placed on the chest, or how dry or moist is the skin of the body, or whether

there is any scaring or irritation on that part of the skin. The arrow from the patient to the interface indicates that the 'values' of the interface are dynamic. For example, a patient could change the location of electrodes during operation, or skin irritations could develop over time if electrodes are worn for long periods of time.

The coupling interface in one way serves as a description of the spatial configuration of the system. We could say that it 'selects' the patient dynamics that are visible to the BSN or 'controls' how patient dynamics are made visible to the BSN based on this configuration. This configuration includes all the factors mentioned above. Using the electrocardiography example, the aim of the sensor is to sense electrical signals generated by the heart to control beating. These signals manifest on different points on the body. The signals that are 'seen' by the sensor depend on the configuration of the electrodes (the placement, tightness of fit and conditions of the skin). For the same signaling pattern in the heart, different configurations will experience different signals. The patient dynamics are the same, but what is presented to the sensor depends on the configuration [41].

A configuration 'value' ($c_{S\leftrightarrow H}$) on the coupling interface is quite complex to describe mathematically and may depend on the particular BSN being described. It would definitely contain a location parameter which could be a three-dimensional quantity (or set of such quantities) specifying a point (or area) on the body. The nature of the coupling, if we are interested in looseness, could also be a three-dimensional parameter related to the amount of deviation in each dimension. The nature of the point of coupling could be represented by function of three-dimensional location input to the a value representing the nature of the point of contact (*e.g.,* amount of skin irritation or moisture). The time evolution is usually considered continuous.

### 5.2.2 The Human-to-Sensing-Subsystem Interface ($H \rightarrow S$)

The patient and the environment produce physical outputs that the BSN is sensitive to. The BSN may not be 'interested' in all these outputs, and some of these outputs may hinder its ability to achieve its goals. For example, electrical interference from other sources (remember that the patient environment is part of $H$) may show up in an electrocardiograph (ECG) and distort the intended signal. Such inputs to the BSN are termed as interfering inputs if they resemble the signal the BSN is trying to actively sense; or modifying inputs if they are signals or energies that the BSN is not actively sensing, but is sensitive to, and hence can affect its operation [114] (*e.g.,* heat affecting electrical circuitry in the BSN). The signals that the BSN is 'interested' in are termed as desired inputs.

In addition to physical outputs to the BSN, the patient might provide some information output (through a physical interface) to the BSN. These could be configuration information like height or weight that allows the BSN to adapt properly to the patient, or responses to queries by the BSN. In either case, the user may have to provide this information because the BSN may not be equipped to sense the information, or sensors for such information are not available. Though physical interaction is still required, it is more convenient to separate what are more informational in nature like configuration information from physical signals.

The physical quantities would usually be modeled as continuous quantities (usually with continuous time). The informational quantities would usually be modeled as discrete quantities (usually with discrete time).

### 5.2.3 The Sensing-Subsystem-to-Human Interface ($S \rightarrow H$)

The BSN may produce physical outputs that the patient is sensitive to: some of these outputs may be intentional energy exposed to the patient to aid in sensing (*e.g.,* light energy used in pulse oximetry), while others may be produced due to the physical nature of the BSN (*e.g.,* heat from electrical circuity or chemicals in batteries or physical packaging of BSN components). Some of these outputs could trigger a reaction by the patient that alters the patients structures around the point of contact (*e.g.,* chemicals in packaging irritating the skin), which would in turn alter the configuration ($\mathscr{C}_{S\leftrightarrow H}$) which would then affect the outputs that the BSN sees ($Y_{H\rightarrow S}$). Others could harm the patient by introducing undesirable side-effect outcomes.

Also, since some of these physical outputs become part of the patient's environment, depending on the configuration, an output created by one BSN component could be an interfering or modifying input seen by another BSN component on the human-to-sensing-subsystem interface ($Y_{H\rightarrow S}$) and affect that component's function. An example of where this occurs is in the case of wireless communication, where components can interfere with each other if multiple components are trying to transmit information at the same time.

In addition, the BSN could provide informational outputs queries to the patient in order elicit information from the patient. Informational outputs could be system status messages so the user knows what mode the BSN is in or whether to charge or replace batteries. Queries could be questions on a survey relevant to the sensing and inference that BSN performs. In addition, we could even stretch the model to include informational outputs like that contained in a user manual on this interface. That information will certainly affect the way the patient interacts with the BSN.

As with the human-to-sensing-subsystem interface, the physical quantities would usually be modeled as continuous quantities (usually with continuous time). The informational quantities would usually be modeled as discrete quantities (usually with discrete time).

### 5.2.4 The Sensing-Subsystem-to-Decision-Making Interface ($S \rightarrow D$)

The goal of the BSN is to produce medically-relevant information for decision-making. These come in the form of digital quantities (discrete quantities with discrete time). In addition, the BSN could also send other information like status information to the decision-making subsystem. We could also stretch the model to include informational outputs like those in a user manual.

Based on the medically-relevant information, the decision-making subsystem instructs the actuation subsystem in what quantities to produce for the patient (on the $A \rightarrow H$ interface). The nature of this medically-relevant information produced by the BSN can be quite complex and since knowledge of this nature is important for identifying hazards, its abstract form is detailed below (with examples). The general forms of this kind of information ($Y_{S \rightarrow D}$) on the interface is first described, then timing of the information is discussed, and both are finally used to describe the abstract form of the BSN output.

#### General Forms of Body Sensor Network Information Output

From the perspective of the decision-making subsystem, the information received by the BSN could be viewed as a stream or signal. The different forms these signals are described below, assuming that the decision-making subsystem is a clinician, in order to refer to a more concrete version of the subsystem. The particular choice decision-making subsystem is inconsequential and all points raised below apply to all choices of these subsystems.

**Sets of values.** The simplest general form of information output from the BSN are sets of values. These are essentially snapshots of information from the patient which are not ordered. For example, a clinician may be interested in the blood pressure readings for a patient over the course of a day in order to establish a minimum, an average and a maximum, but may not necessarily be interested in the time at which these readings are taken or when each reading occurred relative to the others. Another example may be that the clinician may be interested in the lengths of walks that a patient took over a day or a week, but again not in when these occurred or how they are related in time.

**Ordered sequence of values.** The next form of information output from the BSN is an ordered sequence of values. Here, the values are ordered by time, even though their actual occurrence time or the relative time between values are not indicated. A patient may use a device to measure blood pressure in the morning after waking up, sometime in the afternoon, and in the evening. The actual times may not be known, but the order in which blood pressure readings were taken would.

**Time-stamped and relatively-timed sequence of values.** In some cases, timing information about the data is important and hence the BSN may produce time-stamped sequence of values. These time stamps are usually in 'wall clock' time, and are typically used when multiple streams of information from the BSN need to be correlated. In the examples given previously, the wall clock times of all the readings can be logged by the system. When the time stamps are not in wall clock time but in some other time reference like a system clock, or when the time between values is known but their wall clock times are not, we call this a sequence of relatively-timed values. An ECG strip could be an example of such a sequence since it is sampled at a constant rate; the relative times between samples are known even if the actual wall clock times of the samples are not.

**Time-ranged sequence of values.** In some cases where discrete events are being monitored or summary information is being presented, the information may reflect the value over a period of time. This type of information is a time-ranged sequence of values. For example, if the BSN reports that a patient was walking or jogging over a particular period of time (say from 5:30pm to 6:15pm), this particular activity becomes the value for that range of time.

**Hierarchical combinations of signal types.** An information stream could be a hierarchical combination of these forms of information. For example, a BSN may report the sets of ECG samples from detected arrhythmias over a day. This would be the highest level in the hierarchy of the information stream. Each ECG sample is a relatively-timed sequence of values since the information is a waveform sampled at a particular frequency. This comprises the next level of the information stream.

### Timing of Body Sensor Network Information Output

A BSN is typically a software-regulated system and usually has communication networks for coordinating operations between its various components and for communicating with the clinician. Thus, it inevitably exhibits particular temporal behaviors.

Assume that the BSN is tracking a continuous signal produced by the patient as shown in figure 5.3, and that it needs a segment of this signal of time size $t_{seg}$ in order to compute the 'value' of the information. For example, the BSN could be a pulse oximeter that needs to monitor a few milliseconds of the photoplethysmograph in order to compute the heart rate. At time $t_1$, the BSN obtains a segment of the signal comprising data points between $t_0 = t_1 - t_{seg}$ and $t_1$. Since it takes some time for the BSN to compute the 'value' of the segment, the BSN may want to capture the time when the signal for computing the information was acquired. It may have a clock to for doing this and may log the time when it obtains the last point of the segment as $t_2$, where $t_2 - t_1 \geq 0$.

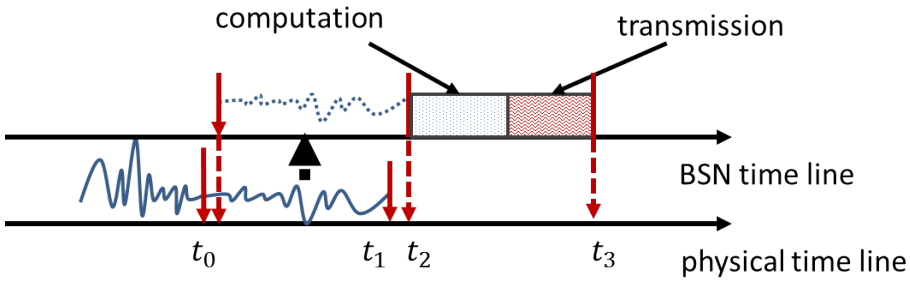



Figure 5.3: Illustration of temporal phenomena in BSNs with respect to sensing, computation, and transmission of information.

Let us assume that the BSN transmits its value to a remote point where the clinician can obtain this information as soon as the information is available at the remote point. The information from BSN will be available at this point at some time $t_3$, where $t_3 > t_1$. When the information is reported,the BSN will assign $t_2$ as the time for that particular value. This time can be termed the *reported observation time* of the particular value. The clinician (representing $D,A$) is first able to obtain this value at $t_3$. This time can be termed the *received time* for that particular value. The time $t_1$ (or $t_0$ if the system logs the time of all points in the segment) can be termed as the *actual time*.

### Abstract Information Model for Body Sensor Networks

As mentioned previously, the model of output from the BSN to the decision-making system ($Y_{S\to D}$) is a stream or signal. Each stream could consist of multiple sub-streams ($Y^i_{S\to D}$). The requirement is that each of sub-streams be homogeneous, in the sense that they must contain the same type of information. Hence, a system that reports activity data correlated with heart rate data may have activity as one sub-stream and the heart rate information as another substream.

An information (sub)stream contains a number of data points. For a substream ($Y^i_{S\to D}$), each data point ($y^i_j$) should be a tuple $y^i_j = (t(y^i_j), v(y^i_j), t_r(y^i_j))$, where $t(y^i_j)$ is the reported observation time, $v(y^i_j)$ is the value, and $t_r(y^i_j)$ is the received time of $y^i_j$. The two different times require some explanation. The reported observation time is the time (usually wall-clock) the BSN claims the event related to the data point occurred (regardless of when the decision-making subsystem is informed about the data point). The received time is the earliest time the data point is available to the decision-making subsystem (*i.e.,* the earliest time it can access the data regardless of when it actually does access the data). Note that $v(y^i_j)$ could itself be a set of data points $y^i_{j,k}$, which have the same tuple form as $y^i_j$ (*i.e.,* they possess the same properties as a data point in a stream). The recursive nature of $v(y^i_j)$ allows us to model hierarchical streams.

It should be noted that the term "value" is used loosely as a value could be as complicated as an image taken by the BSN (in case of a mobile ultrasound, for example). The form of the reported observation time ($t(y^i_j)$) determines the stream type. Table 5.1 defines different forms of $t(y^i_j)$ and the stream types.

Table 5.1: Different forms of reported observation times and associated stream types

| Form (Notation) | Meaning | Stream Type |
|---|---|---|
| $t^i_j$ | time stamp or relative time | time-stamped or relatively-timed |
| $[t^i_{j,0}, t^i_{j,f}]$ | time range | time-ranged |
| $k^i_j$ | ordering index | ordered |
| None or tag | no reported time (but labeled) | set |

Figure 5.4 shows an example output stream. This stream, $Y^{ecg}_{S\to D}$, contains sets of data points $\{y^{ecg}_i \ldots y^{ecg}_n\}$, each of which is an ECG strip. Each ECG strip contains a sequence of relatively-timed data points (voltage values sampled at 250 Hz). Formally, $y^{ecg}_i = (i, v(y^{ecg}_i), t_r(y^{ecg}_i))$, where $i$ is the tag for the strip, and $v(y^{ecg}_i)$ is also a sequence of

relatively-timed data points $\{y_{i,1}^{ecg} \ldots y_{i,m}^{ecg}\}$, such that $y_{i,j}^{ecg} = ((i,j), v(y_{i,j}^{ecg}), t_r(y_{i,j}^{ecg}))$, $t_{j+1}t_j \approx$ 4ms, and $v(y_{i,j}^{ecg})$ is a digital value representing the electrical potential measured at the patients skin surface. If the BSN is assumed to be streaming the ECG data, then the received time of each sample can be assumed to be a small delay from the actual time when the value was measured (*i.e.*, , $t_r(y_{i,j}^{ecg}) - t_0(y_{i,j}^{ecg}) < \varepsilon$, where $t_0(y_{i,j}^{ecg})$ is the actual time of the sample $y_{i,j}^{ecg}$).

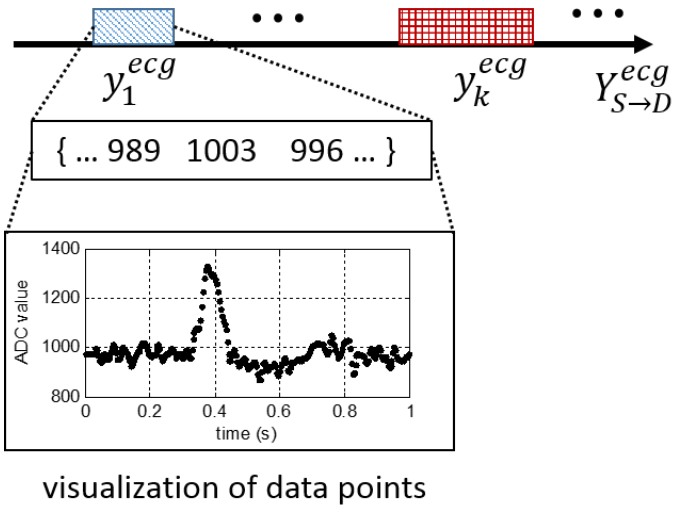



Figure 5.4: An example of a stream from BSN showing the hierarchy of substreams

### 5.2.5 The Decision-Making-to-Sensing-Subsystem Interface ($D \to S$)

The BSN may also receive information ($Y_{D\to S}$) from the decision-making subsystem during operation. These could be configuration commands in order to change BSN operation modes or query commands to elicit information from the BSN. They could also be acknowledgments of receipt of information sent by the BSN. The information on this interface would usually be modeled as digital (discrete values and discrete time).

## 5.3 The Generic Set of Hazards for Body Sensor Networks

As mentioned in chapter 2, a hazard is a system state or condition which with a worst case set of environmental conditions could lead to an accident or mishap. The mishaps we are interested in are unacceptable functional or side-effect outcomes for the patient. A BSN could create hazards and directly harm the patient through the sensing-subsystem-to-human interface ($S \rightarrow H$), or indirectly harm the patient through the sensing-subsystem-to-decision-making interface ($S \rightarrow D$) by causing the decision-making subsystem to make a hazardous decision. Below we identify the generic set of hazards for the BSN based on the nature of these interfaces. The hazards are expressed in terms of BSN behaviors, particularly, the nature of the trajectories of signals on the interfaces.

Note that the other interfaces not mentioned can contribute to hazards, however, we consider their contributions as causal factors since they are not linked directly to the intrinsic behavior of the BSN. Remember that the intrinsic behavior dictates how a system (in this case BSN) responds to inputs to produce outputs (or how it may autonomously produce outputs). Since the BSN is the subsystem of concern, its intrinsic behavior influences what happen at the output interfaces, and since hazards are typically linked to the intrinsic behavior of the system, then its output interfaces are where the hazards should be identified. There are some interactions on the output interfaces of the BSN that are considered more as part of the causal factors than hazards since their effects show up in the hazards identified below and these are discussed in the section on causal factors (5.4.2)

### 5.3.1 Physical Interaction Hazards

For any of the physical quantities ($i$) produced that the human is sensitive to, there is usually a threshold ($\overline{Y}^{i}_{S\rightarrow H}$) below which no harm is caused and above which if sustained for a period will result in significant harm. Usually, if a quantity goes above the threshold for a brief period of time but does not result in harm, there is a 'cool-off' period that must be allowed for the body to be rid of the temporary effects of reacting to the quantity going above the threshold before this can occur again. A formal statement of the phenomena above is represented in the hybrid system model shown in figure 5.5.

When the physical quantities are below the no-hazard threshold ($(Y^{i}_{S\rightarrow H}(t) \leq \overline{Y}^{i}_{S\rightarrow H})$), no hazard is present. When the quantity goes above the threshold, its effect begins to accumulate resulting in a potentially hazardous situation. This is accumulation is captured by the positive derivative of the accumulation variable $b^{i}_{S\rightarrow H}(t)$. The rate of accumulation

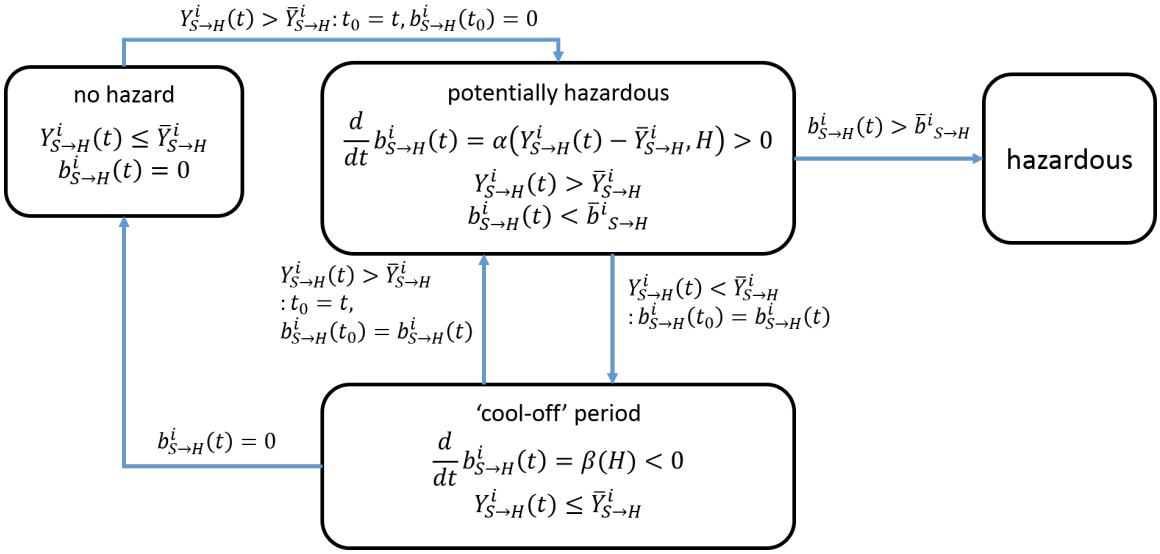



Figure 5.5: Hybrid system model of physical interaction hazard phenomena.

depends on how far above the no-hazard threshold the value of the physical is, and on parameter of the patient as shown in the equation. This allows short but significant deviations above the no-hazard threshold and long but less significant deviations to be both accounted for. When this value accumulates beyond a certain value (the hazard threshold $\bar{b}^i_{S\to H}$), harm can occur and hence the situation becomes hazardous.

If the physical quantity drops below the no-hazard threshold before its effect is accumulated beyond the hazard threshold, the system enters a'cool-off' period where the accumulated effects dissipate (indicated by the negative derivative of the accumulation variable). If this value completely dissipates, the system goes back to a no hazard state. If the physical quantity goes above the no-hazard threshold before the accumulated effects have dissipates (*i.e.,* before the cool-off period has been satisfied), the effects begin to accumulate again, starting from where the accumulation variable was before the physical quantity went above the threshold.

**Example 5.1:** Heating of the Skin by BSN Components

A concrete example of when this hazard could occur is with a BSN component that generates heat. One component that can generate significant amounts of heat is a gas sensor like an ozone sensor (the MICS-2610 sensor from e2v Technologies [34],

for example, needs its sensing resistor kept at around 430°Cduring operation). Ozone sensing can be crucial for management of care for those with asthma [108].

Based on information on fire and heat dynamics from the National Institute of Standards and Technology [74], the physical hazard from heating of the ozone sensor can be represented by the hybrid system model in figure 5.5 as follows.

Let $Y^i_{S\to H}(t)$ be the amount of heat that is received on the skin of the user, and $b^i_{S\to H}(t)$ be the amount of heat accumulated above the normal body temperature (represented by $\overline{Y}^i_{S\to H}$) of 37°C. In general, so long as the heat felt by the skin from the sensor is below the normal body temperature (or less than a degree above it), we should be be in the no hazard state. However, once the component produces heat that is felt significantly above 37°C, the skin starts to heat up according to one (or more) standard heat transfer equations for conduction, convection, or radiation, depending which mechanisms factor into the heat felt at the skin. All the equations are rate of heat accumulation equations and hence will be of the general form

$$\frac{d}{dt}b^i_{S\to H}(t) = \alpha(Y^i_{S\to H}(t) - \overline{Y}^i_{S\to H}, H) \tag{5.1}$$

based on the difference between the temperature at the skin (captured in $H$) and the heat from the sensor as well as properties of the skin (also captured in $H$).

In general, above a temperature of 44°C, the skin feels pain from heat and above 48°Cfirst degree burns occur. The accumulation threshold for hazards ($\overline{b}^i_{S\to H}$) can therefore be set at 44°C. If the heating from the sensor stops before this temperature is reached, the skin will cool down according to the standard cooling process. If it cools to normal body temperature, then we are back in the no-hazard state. If the sensor starts to heat again, the return to the potentially-hazardous state where the skin heats up, but now staring from temperature the skin was at before the heating started again.

This heating and cooling of the skin can go (in theory) on as long as we do not enter the hazard state.

Physical interaction hazards are the more obvious hazards of the BSN since they are based on direct interactions. Issues of biocompatibility, for example, are standard issues to consider for most medical devices that come into contact with the human body [117]. In addition, explorations of BSN safety are typically focused on direct physical interaction hazards. For example, Banerjee *et al.* applied formal verification to assess thermal safety

associated with the interaction between the BSN and the patient [11], and De Santis *et al.* used modeling techniques to study the potential risks of ultra-wide band (UWB) radios for the patients [31]. The more subtle aspect of BSN safety is the indirect effects through the sensing-subsystem-to-decision-making interface discussed below.

### 5.3.2 Information Quantity Hazards

Such hazards arise when the amount of data points produced by the sensing process deviates from the expectation. This occurs when events go undetected or unreported by the BSN or when the BSN produces spurious information. Recall the meal sensing example in example 4.13, where we established that if the meal sensing acted in a way that allowed more than one meal (out of three that were actually taken) to go unreported then this resulted in unacceptable risk. The scenario where more than one meal goes unreported is an example of an information quantity hazard. In this case, there is less information than expected. If the meal sensing system spuriously reports a meal that was actually not taken (a case that was not explored in that example), and if this resulted in unacceptable risk, then that would be a case of an information quantity hazard where there was more information than expected.

Information quantity hazards can be defined formally using the behavior space and acceptable behavior region approach established in the discussion of safety-guided in section 4.4.3. Let us assume that for each piece of information that the BSN generates, there is a parameter ($\lambda_S^{q,i}$ where $i$ is the 'tag' for the specific piece of information) that describes whether information goes unreported (for whatever reason) and/or spurious information can be produced. One way to think of this parameter is to think in terms of the actual information available ($|\overline{Y}^i_{S\rightarrow D}|$) and the information that is reported to the decision-making subsystem ($|Y^i_{S\rightarrow D}|$) in a finite period of time of the system.

If we look at the ratio of reported information to the actual information then a value of 1 would indicate that amount of reported information is correct, a value less than 1 would indicate underreporting of information, and a value greater than 1 would indicate reporting of spurious information. The parameter $\lambda_S^{q,i}$ could be a single value greater than 0 or it could represent the range of possible values the BSN might take (especially if the BSN can both underreport or produce spurious values and this behavior is variable).

Based on the definition of this parameter, we can establish a behavior criteria ($\overline{\lambda}_S^{q,i}$) which describes how much unreported and/or spurious information is acceptable. If a BSN consistently underreports or produces spurious information in a way that does not satisfy

this criteria (*i.e.,* has a $\lambda_S^{q,i}$ value that does not satisfy the definition of $\overline{\lambda}_S^{q,i}$), then it creates information quantity hazards.

**Example 5.2:** Information Quantity Hazards in Activity Tracking

Let us assume that the BSN must track and report when a person takes short (less than 5 minute) walk at any point in the day. If we assume that the person internally has a symbolic state variable ($x_{act}$) representing the particular activity the person is currently undertaking, then the job of the BSN then is to infer the each time this variable takes on the value representing walking for less than 5 minutes.

If the BSN works perfectly with respect to this inference, it will produce a stream $|\overline{Y}_{S\to D}^{walk}|$ of data points each representing an actual short walk that the person took. An actual BSN would produce a stream $|Y_{S\to D}^{walk}|$. Let us assume that the information quantity parameter ($\lambda_S^{q,walk}$) is the ratio of reported information to the actual information. Let us further assume that through the requirements process in the safety-guided design, we have established that so long as the number of short walks report is within 10% of the actual value (above or below), the overall system results in acceptable risk. Hence our criteria is

$$\overline{\lambda}_S^{q,walk} : 0.9 \leq \lambda_S^{q,walk} \leq 1.1 \tag{5.2}$$

Let us also assume that BSNs designed for this purpose are characterized by the range of values $\lambda_S^{q,walk}$ can take (*i.e.,* it is capable of both underreporting or generate spurious walking events but up to a certain amount), then any BSN whose $\lambda_S^{q,walk}$ lies inside the range of the criteria does not result in information quantity hazards. If any part of $\lambda_S^{q,walk}$ for a BSN however lies outside the range, then that BSN can produce information quantity hazards.

### 5.3.3 Received Time Hazards

These occur when a (set of) data point(s) in a stream is reported too late. This is when there is a significant delay between the time that data point is generated and when the decision-making subsystem can first access the data point, and this delay results in unacceptable risk. One way to think of this hazard is that once the particular data point is generated, there is a time within which a decision based on that data point must be made, and if the

data point is received after this time period has passed, then the information it provides is no longer actionable.

We can define this hazard formally as follows. For each data point ($y^i_{S\to D}$) in a stream, there is a time when the point is actually generated in the system ($t_g(y^i_{S\to D})$). Assuming that the time the data point is generated is when it should be, there is a messaging window ($\lambda^{\delta_{t_r},i}_{S\to D}$) between when the point is generated and the the decision-making subsystem learns about the data point. A received time hazard occurs when this window exceeds a limit ($\overline{\lambda}^{\delta_{t_r},i}_{S\to D}$), which when exceeded prevents the decision-making subsystem from responding appropriately to the information contained in the data point to prevent mishaps. That is

$$\lambda^{\delta_{t_r},i}_{S\to D} = t_r(y^i_{S\to D}) - t_g(y^i_{S\to D}) > \overline{\lambda}^{\delta_{t_r},i}_{S\to D} \Rightarrow \text{hazard} \tag{5.3}$$

where the times ($t_r(y^i_{S\to D})$ and $t_g(y^i_{S\to D})$) are assumed to be wall clock times.

**Example 5.3:** Delayed Hypoglycemia Alarm for Nighttime Glucose Monitoring

Let's assume that the BSN is a night-time glucose monitoring system for children, which alerts the parents (in another room in the house) if there is risk of hypoglycemia. If when hypoglycemia occurs, the system detects it immediately, but due to network issues, it is unable to get alert out to the parents in time for them to intervene to prevent an emergency room visit, then this would be received time hazard.

### 5.3.4 Reported Observation Time Hazards

These occur when the reported observation time for a data point in a stream differs from the actual time of occurrence of the event or value such that it results in unacceptable risk. Note the subtle difference between the reported observation time hazard and the received time hazard. In the received time hazard, it is assumed the the data point is generated by the BSN very close to when the event related to the data point occurred, but this data point is made known to the decision-making subsystem much later.

In the reported observation time hazard, it is assumed that the decision-making subsystem gets to know about the data point within a reasonable time from when it is generated, but the issue is that the BSN either anticipates the event much earlier than it occurs of it learns of the event much later than its occurrence and assumes that when it first learns about the event is when it occurs.

One way to think about it is that reported observation times are concerned with timing issues between phenomena occurring in the human dynamical system to the sensors ability to infer these phenomena, and received time hazards are concerned with delays of information transfer from the BSN to the decision-making subsystem.

We can define the reported observation time formally as follows. For each correctly-generated data point ($y^i_{S\to D}$) in a stream, there is a time when event that prompted the generation of that data point occurred ($t_0(y^i_{S\to D})$). There usually a difference between ($\lambda^{\delta_t,i}_{S\to D} = |t(y^i_{S\to D}) - t_0(y^i_{S\to D})|$) this time and the time the BSN claims the event occurred. When this difference exceeds a threshold ($\overline{\lambda}^{\delta,i}_{H\to D}$), it affects decision-making in a way that results in unacceptable risk. That is

$$\lambda^{\delta_t,i}_{S\to D} = |t(y^i_{S\to D}) - t_0(y^i_{S\to D})| > \overline{\lambda}^{\delta_t,i}_{S\to D} \Rightarrow \text{hazard} \tag{5.4}$$

where the times ($t(y^i_{S\to D})$ and $t_0(y^i_{S\to D})$) are assumed to be wall clock times.

Note that this hazard is not a real-time requirement. For retrospective analysis, it could cause problems because if both the BSN data and actions taken by the decision-making based on this data are reviewed. The decision-making actions may be evaluated incorrectly since the temporal relationship between the events and the responses may not be captured properly. Also, if multiple information streams are being evaluated retrospectively, reported observation hazards indicate that the streams are not synchronized properly and hence temporal relationships between different events across streams may not be captured properly as well.

**Example 5.4:** Reported Observation Time Hazards in Sensing for the Artificial Pancreas

**Meal Sensing** In the meal sensing example (4.13), it was possible for meals to be reported (as much as an hour) before or after they were actually taken. In that example, what was actually being reported was a 'meal-warning' indicating that a meal was about to be taken soon or had been taken not long ago (usually within 20 minutes of when it was reported). We saw that when the meal is taken more than 20 minutes before it was reported, this resulted in increased population level risk.

In addition, if we were doing a retrospective analysis to improve our glucose management strategy, mismatch between when meals were reported and when they were

taken would cause us to arrive at wrong conclusions about the performance of our management strategy and its response to meals.

**Blood Glucose Level Sensing** Breton and Kovatchev showed that the value that a continuous glucose monitor reports is the blood glucose value about 10 to 15 minutes before the sensor reports it [18]. This is because these sensors do not sense the glucose in the blood directly, but rather they sense the glucose in the interstatum to infer the blood glucose level. There is however, a delay of diffusion of glucose from the blood to the interstatum, which means the sensor is physically operating on delayed information. This is why glucose control algorithms [29] tend to have a predictor component which essentially estimates the current glucose value and next value would be based on past values so that it can act on at least an estimate of the current value.

### 5.3.5 Value Hazards

Value hazards occur when the values reported by the BSN cause the decision-making subsystem to infer differently from what actually happened at the human dynamical system and to take actions that result in acceptable risk. The values presented by the BSN could be information with complicated structures.

To understand value hazards, we can think of the BSN as a translator. In one sense, it translates trajectories observed at the human-to-sensing-subsystem interface to trajectories that are meaningful to the decision-making subsystem. If we think of each interface ($H \rightarrow S$ or $S \rightarrow D$) as having its own 'language', the BSN translates sentences (trajectories) from the language spoken by the $H$ to the rest of the world to sentences in language spoken by $D$. In turn we can think of $A$ as speaking a language understood by $H$ (not necessarily the same one spoken to $S$) in order to affect its actions, and $A$ as a translator from $D$ to $H$. If we assume that $A$ is a perfect at its job, then value hazards occur when the BSN misinterprets (for whatever reason) what it 'hears' from $H$ and the result of the misinterpretation is a significant misunderstanding at $D$ of what $H$ is trying to say such that $D$'s actions result in harm to $H$.

Using the above, we can define for each stream ($i$) the BSN produces for $D$, a parameter ($\lambda^{v,i}_{S\rightarrow D}$) that characterizes this susceptibility to misinterpretation by the BSN. A BSN produces value hazards if its susceptibility is not within the tolerable range ($\overline{\lambda}^{v,i}_{S\rightarrow D}$).

**Example 5.5:** Value Hazards in Artificial Pancreas Sensing

**Meal Sensing** In the meal sensing example, we showed that carbohydrate content of meals were underestimated, then this would result in population-level risk. This was independent of when the meal was reported.

**Blood Glucose Level Sensing** Work by Boyd and Burns [17] and Patek *et al.* [52] investigated the difference in values between the blood glucose value and what the glucose monitor reports and how these could potentially affect decision-making. They looked in particular at intrinsic noise behavior.

Another issue in continuous glucose monitors are what typically termed pressure-induced sensor artifacts [14] which arise due physical interactions of the sensing element with the body. This create deviations in values that are not correlated with changes in the actual blood glucose and can be a problem especially for the predictors in controllers which rely on Kalman filtering which in turn rely analysis of past trajectories. These can result in inappropriate insulin infusion.

## 5.4 Utility of the Model and Set of Hazards

There are three main things the model and the set of hazards allow us to do. The third is the ability to characterize BSNs in blackbox manner based on the hazards in order to identify acceptable behaviors for a safety-guided design process. The second is the ability to identify causal factors for the hazards based on knowledge of the nature interactions on the various interfaces and behaviors of the components in the model. The second is the ability is to develop a (non-exhaustive) general set of points to keep in mind to guide the design process based on the hazards and knowledge of the nature of behaviors of and interactions between components in the model. These points are discussed below.

### 5.4.1 Black-Box Characterization of Body Sensor Networks for Safety Guided Design

In Section 4.4.3 when safety-guided design was discussed, the goal was to identify black-box behaviors of the subsystem under consideration in order to identify acceptable behav-

iors that would factor into requirements for more detailed designs. The idea was that these black-box behaviors can be specified as expectations on interactions between the subsystem of interest and other subsystems in the health management scenario (*i.e.,* interface requirements).

In the example (4.13), an interface was developed between meal sensing and decision-making in order to identify the acceptable behaviors for the meal sensing. We, however, did not derive a general systematic approach for deciding what parameters must be include as part of the interface, and focused solely on issues at the sensing-subsystem-to-decision-making interface. This was because the behaviors at this interface are most subtle in the way they affect safety. They also happen to be the behaviors that are most amenable to black-box characterization because it is focused on the conceptual goals of the BSN and this requires no detailed knowledge of the implementation. The issues at the sensing-subsystem-to-human interface tend to be more implementation-specific. This section focuses on this general systematic approach for identifying the interface parameters.

The key is in the general set of hazard we just identified. For each hazard, we actually developed a parameter (or behavior) at the interface and then stated that there are some values of these parameters that correspond that will result in hazards and others that will not. This reminiscent of the way we discussed acceptable behaviors in the safety-guided design approach in section 4.4.3. There, we stated that behaviors could be described in terms of parameters at the interface. The requirements process then derived a limit on the range of values these parameters can take (or variability that these parameters can exhibit). The idea was that if behaviors are outside this range, then unacceptable risk occurs.

Recall that hazards are those conditions that can lead to unacceptable risk. This means that the acceptable behavior region in each parameter dimension essentially corresponds to a hazard criteria. Hence, the black-box interface parameters for the BSN should correspond to those parameters that were developed in order to define the hazards. The received time and reported observation time parameters were described properly in the development of their hazards. The others, however, need a bit more elaboration.

### Physical Interaction Behavior

This black-box behavior ($\lambda^{i}_{S\rightarrow H}$)for each physical quantity of interest can be characterized by a hybrid system similar to the one used to define the physical interaction hazard. This is shown in Figure 5.6.

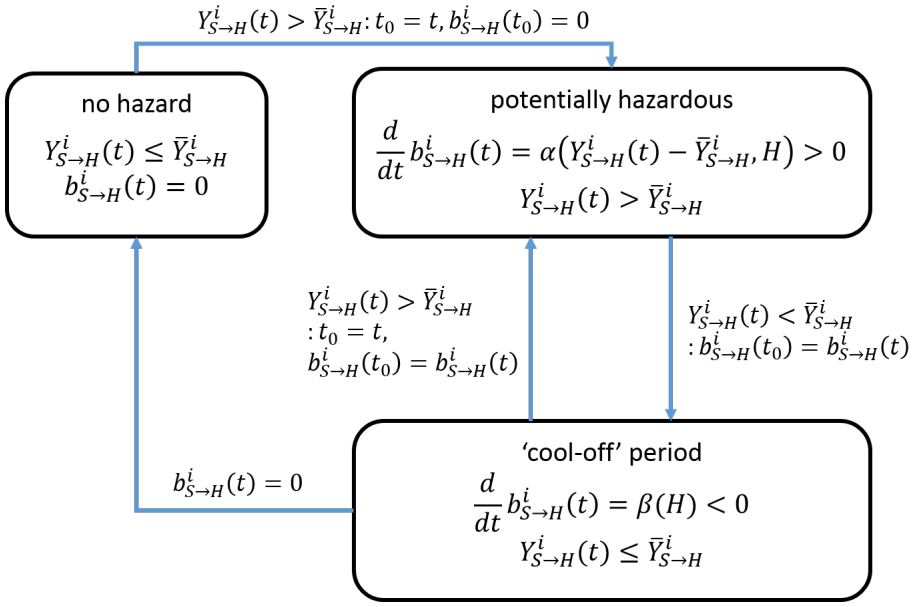



Figure 5.6: Hybrid system model of physical interaction behavior of BSN related to hazards.

Note that the behavior shown here does not include the hazard state that was present in figure 5.5. This is because we are only interested in the BSN's behavior, either for the purposes of determining what the limit for transitioning into the hazard state should be, or for assessing if this behavior would result in a transition into that state (if we already know what this limit is).

For a particular physical quantity, the transition behaviors (the likelihood that it transitions from one state to another or stays in that state) might vary from one BSN design to the next and this transition behavior is what characterizes the BSN with respect to that particular quantity. If these transitions are stochastic in nature, then they could be characterized by a transition matrix. If the transitions are stochastic, then acceptable behaviors could be described as properties of the transition matrix.

### Information Quantity Behavior

When defining the hazards, we glossed over what actually goes into the parameter, and used a simple example for our developments. The particular way to define this parameter for a specific (black-box version of a) BSN depends on the nature of the information that it

must produce. Also because of the dynamic nature of information that the BSNs in general produce, the parameter should be thought of more like a transfer function, or a mapping of sets of trajectories to another set of trajectories.

For a given set of trajectories produced by the patient on the $H \rightarrow S$ interface, there is some embedded information ($\overline{Y}^{i}_{S\rightarrow D}$) that the decision-making subsystem expects to receive. The BSN can be thought of as applying a function to this expected information to produce the reported information ($Y^{i}_{S\rightarrow D}$). The information quantity parameter characterizes how will this function preserves the amount of information in $\overline{Y}^{i}_{S\rightarrow D}$.

**Example 5.6:** Information Quantity Paramter for Meal Sensing

Recall that in the meal sensing example (4.13) we looked at the possibility of meals going unreported. There, we assumed that there were always three meals in the day and that the BSN could only miss reporting a meal (*i.e.,* it does not produce spurious meal information). In defining the requirements, we came up with a meal report indicator

$$\begin{bmatrix} \omega_B \\ \omega_L \\ \omega_S \end{bmatrix} : \omega_i = \begin{cases} 1 & \text{unreported meal} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

In that case, we were not concerned about which specific meals went unreported so our information quality parameter ($\lambda^{q,meal}_{S\rightarrow D}$) would be the number of unreported meals

$$\lambda^{q,meal}_{S\rightarrow D} = \sum_{i\in\{B,L,S\}} \omega_i \tag{5.6}$$

If the number of meals in a day is variable, then the information quality parameter ($\lambda^{q,meal}_{S\rightarrow D}$) could represent the likelihood that any meal goes unreported. If we are also concerned about spurious meals, the parameter could be a conditional probability distribution indicating the likelihood that a certain number of meals are reported given that a certain number of meals were actually taken. If $Y^{meal}_{S\rightarrow D}$ represents a meal trajectory (*i.e.,* a sequence of meals), then this would be

$$\lambda^{q,meal}_{S\rightarrow D} = p(Y^{meal}_{S\rightarrow D} | \overline{Y}^{meal}_{S\rightarrow D}) \tag{5.7}$$

### Value Parameter

This parameter must be agnostic to timing and information quantity issues. It must only describe how values differ from what might be expected. We must therefore make a few assumptions.

First, recall that value hazards are related to this idea of the susceptibility of BSN to misinterpret information provided to it in the language of the patient when translating it to the language understood by the decision-making subsystem (using translation as analogy for the role of the BSN). Let's assume that for each possible sentence (finite segment of a continuous or discrete signal) from the patient ($Y_{H\rightarrow S}$), there is a perfect translation into a sentence (a discrete signal), in terms of values, in the language understood by the decision-making ($\overline{Y}_{S\rightarrow D}$). The translation could be wrong timing-wise, but this is not the concern of the value parameter. This is necessary because all the parameters are described in terms of some deviation from the ideal.

Given the above, one way to model the way the BSN misinterpret information is to assume that it is internally composed of a perfect translator and an imperfect messenger. Hence the BSN always (but not in reality) starts with the perfect translation of what the patient tells it, but the imperfect messenger mishandles the message. Let's places some requirements on the messenger. First it must preserve the order of the values in the sequence (discrete signal) it receives from the perfect translator, and must not affect the value of the received time (in order to preserve timing). Second, the number of values in the sequence must be preserved in order not to affect information quantity. With this, the messenger can be defined as an operation on this sequence, a mapping from a domain of the values of the sequence to a range which is the same as the domain of the sequence. The value parameter then becomes a (set of) parameter(s) of this operation.

One example of mathematical operations that model the messenger is a discrete filters. A discrete filter when applied to a sequence produces another sequence of the same size as the input sequence. A filter can be parameterized in a number of ways, one being specifying the tap coefficients, and the other being specifying the type, cut off frequency, and what is known as a Q parameter. Either parameterization would correspond to the value parameter for the BSN if it behaves like a filter. Notice that the discrete filter usually requires a number of points in the input sequence to produce one value in the output sequence. It however preserves order of values in the sequence.

Another example of a mathematical operation that models the messenger is a random noise process. This process adds a random value to each value in the input sequence. The

way it picks the random value can be described by a simple probability distribution or a stochastic process. The parameters of either approach become the value parameter.

**Example 5.7:** Value Paramter for Meal Sensing

Recall that in the meal sensing example, there was a meal factor that determined how different the reported carbohydrate content was from what the patient actually took. This meal factor only operated on the meal values regardless of the timing of the meals. There were 4 possible values for each meal (4/7x, 1x, 1.5x and 2x)for the factor. We considered the actual meal factor to be the 3-dimensional vector of factors for each meal. In the example we assumed that the choice of any factor (vector) was equally likely which would make the value parameter the parameters of a noise process given by a discrete uniform distribution with the possible values being the $4^3 = 64$ possible values of the vector. Once we identified the hazards, we actually put a limit of which vectors were allowable, essentially changing the nature of the noise process.

### 5.4.2 Potential Causal Factors for Body Sensor Network Hazards

To discuss potential causal factors for some of the hazards, it is important to understand the general internal structure of the BSN. Figure 5.7 shows a version of the generic BSN model with internal structure detailed.

A BSN could be made up of a collection of physically separate subsystems (such as $S_i$, $S_j$, $S_k$, and $S_l$). Some of these BSN subsystems may be located on or close to the body (like $S_i$, $S_j$, and $S_k$) and others located away from the body (like $S_l$). These sub-processes interact with each other and the decision-making subsystem through communication channels (either communication networks or storage media).

Each sensing sub-process ($S_{(*)}$) may contain a number of components, and must at least contain some form of computational element ($S_{SP}$), which is connected to a communication channel to interact with other subsystems or the decision-making subsystem. BSN subsystems that must interface directly with the human process H (like $S_i$, $S_j$, and $S_k$) would require a transducer ($S_{\leftarrow}$) for converting physical quantities produced by $H$ to voltages and, if necessary, an "stimulator" ($S_{\rightarrow}$) for producing energy output that affects $H$ to aid in sensing. A BSN subsystem may also have a (potentially) limited energy source

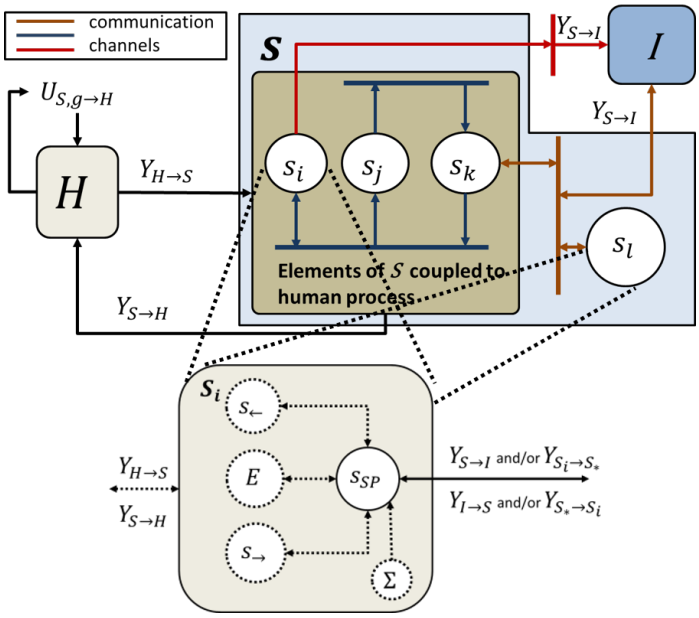



Figure 5.7: The schematic of the generic body sensor network model with the internal structure shown.

($E$) if it is mobile. A sub-process may exist within a computational environment ($\Sigma$) if it shares computational resources with other processes that are not part of the BSN.

### Physical Interaction Hazards

Physical interaction hazards can be caused by inappropriate limitations on the BSNs ability to produce certain quantities like heat and electromagnetic radiation. They could also be caused by material used on the parts that come into contact with the patient. The patient could also damage or introduce chemicals to the physical structure that cause interaction with the skin or other tissue that would not otherwise not occur. The patient could also place BSN components in locations where they should be and cause problems.

### Information Quantity Hazards

Information quantity hazards are typically caused by limitations in detection algorithms in the BSN. Either they are not robust to interfering or modifying inputs or they have limited detection capabilities even under ideal conditions due to poor design or limited resources. The BSN may go into a power-saving mode that reduces its ability to effectively detect the

events of concern. In addition, network issues can cause information to be dropped along the way, and if reliability mechanisms are not available, this could result in missing crucial information.

#### Received Time Hazards

These are usually due to network delays, especially in the case where the decision-making subsystem is further away from the patient geographically. Delays could be malicious or unintentional. They could also be caused by software problems if concurrent processes infer with each other.

#### Reported Observation Time Hazards

These can occur if detection capabilities require too much time to decide if an event occurred or not. In addition, if there are poor time-synchronization mechanisms between the BSN and the decision-making as well as between BSN components, then timestamps can be misinterpreted. For BSNs where the user must produce the inputs necessary for the BSN to deliver information to the decision-making subsystem, the user may forget to put in an input on time and forget a particular time when the event occurred.

#### Value Hazards

These could also be due to the limited processing capabilities and robustness of the BSN. In addition, if configuration or calibration information is required, then this could be due to misconfiguration either by the patient or the decision-making, or by a malicious adversary. If BSN instructions in the manual or provided during operation are not clear, the user could misconfigure the system. If the BSN is incapable of detecting certain misconfigurations, this could result in the hazard.

### 5.4.3 Some Points to Consider for Body Sensor Network Safety

Based on the above, the following points must be kept in mind. They phrased as questions.

#### Intent and Interactions

It is important to understand what the goal of the BSN is. What information does it provide for decision-making? How is this information (usually related to a health metric) defined?

What interactions are necessary to aid in gathering this information? How can other interactions interfere with gathering this information? Is the BSN always worn or only used some of the time? Must the user provide information? What if this information is not provided or not provided on time? What does the BSN assume about its location, if any? What are the issues if these assumptions do not hold?

#### Configuration

Since misconfigurations result in hazards it is important to understand whether the BSN is designed an how this configuration works. Who is allowed to configure the BSN? Does configuration happen only before starting operation, or can it be reconfigured during operation? Are the instruction for configuring the BSN understandable to the intended user population? What are the limits to the BSN detecting misconfigurations, and what are the risks given this limits?

#### Operational Environment

Where would one expect the user to use the BSN? What happens if the user goes to unexpected places with unexpected conditions. Are there any assumed locations on the body? How is the BSN component location enforced. If parts of the BSN exist in a shared computational environment, what are the effects of this environment.

## 5.5 Relation to Body Sensor Network Analysis Work

The are two main aspects to BSN analysis, the issues explored and the models used to explore those issues. Our main issue of concern is safety.

### 5.5.1 Safety

Much of the work in BSN safety has been focused on specific issues. De Santis *et al.* [31] focus on safety issues with ultra-wide band radiation from communication radios, which is a physical interaction issue. Armenti *et al.* and Zegiel *et al.* [4, 121] focused on the potential for information provided by the BSN to mislead a decision-maker into taking an inappropriate action, which is a value hazard issue. Banerjee *et al.* focused on the physical issue of heat as well as interference between components of the BSN [11]. That same work

also looked at verification approaches for BSN and in other work they explored synthesis of correct designs once they have been verified [12]. As mentioned previously, Patek *et al.* [52] looked at acceptable behaviors for continuous glucose monitors used for decision-making by the patient, and Boyd and Burns [17] explored the effects of noise behaviors on clinical decision-making.

The work presented here provides a framework within which all these issue can be considered. It also considers the complexity of the nature of information provided by the BSN. The main difference is in perspective. Much of the previous work has focused on design and issues for specific designs, which means there is limited concern for general tools. We are however concerned with design an regulation issues and are motivated by more general considerations.

### 5.5.2 Modeling

Past BSN modeling has focused more on performance than safety, looking at issues like quality-of-service [107] or energy issues [15, 112]. These are related to safety, but these works do not explore them from that perspective. Again, they also look at specific issues. Banerjee *et al.* [11] are the closest to the work presented here attempting to provide a general model and tool, but the safety issues explored are limited to physical issues. In addition, they, like others, take the hazards as given any specific situation, while this work identifies a general set of hazards, and shows how to develop them for specific cases, providing a framework for reasoning more generally about BSN hazards.

## Summary

This chapter used insights from chapter 4 to develop a model for reasoning about patient safety for BSNs. An important aspect is identifying hazards and causal factors. The hazards depend on appropriate black-box characterization fo the BSN. Once a particular BSN is characterized properly, the safety-guided design approach detailed in section 4.4.3 can be used to determine acceptable behaviors, which then become the hazard criteria. With the generic BSN model, one can then consider how their particular BSN concept could behave in ways that violate the hazard criteria using the knowledge of the general behavior of BSN components described by the model, the causal factors, and the points to consider as starting points.

# Chapter 6

# Implications for Body Sensor Network Design Tools

> The difficulty of the design problem often resides in predicting how an assemblage of such components will behave.
>
> Herbert A. Simon

## Chapter Overview

This chapter demonstrates another aspect of the utility of the model of patient by looking at its implications for body sensor network design tools. In particular, it focuses on the need to explore the outcomes of physical interaction between the body and the sensors on it while accounting for the spatio-temporal nature of the human dynamical system. It introduces some motivating applications and the basic requirements for tools with features that enable the design explorations related to spatio-temporal issues. It then discusses a proof-of-concept simulation tool (still under development) which focuses on inertial sensing and wireless communication whose current instantiation embodies these basic requirements. It discusses the potential to aid in the reasoning about safety of body sensor networks in the context of the motivating applications.*

*The initial ideas for BodySim, the proof-of-concept simulation tool described here were presented in a short paper at the International Conference on Body Area Networks [6] and tool demonstration and the ACM

## 6.1 Introduction

Modeling and simulation play important roles in engineering research and design. The semiconductor industry, for example, is heavily dependent on modeling and simulation tools for integrated circuit design and fabrication. These techniques are especially helpful in the early phases where limited detail is available about the design and where design changes are less costly. High-fidelity models can be employed at the verification and validation stages to complement testing. Modeling and simulation are also important research tools for understanding complex phenomena [95].

In design, modeling and simulation typically aid in understanding the relationship between the system being designed and the environment in which the system is intended to operate. Understanding this system-environment relationship is particularly important for BSNs. In Chapter 5, one of the main causal factors for hazards identified was the interaction between a body sensor network subsystem on the and its environment (including the patient and other subsystems). This is because many sensing modalities are directly driven by the dynamics of the user wearing the BSN. Also, the human body itself is a non-uniform environment with many variables of interest exhibiting both spatial and temporal dynamics for any given user state and activity. This makes issues like the effect of sensor location on system behavior important ones to explore. In addition, the behavior of the user and the characteristics of the environment they are in affect other aspects of the BSN, particularly wireless communication.

Previous BSN modeling efforts have typically concentrated on the BSN components (both software and hardware). Examples include exploring particular performance properties like energy consumption [15] and communication quality of service [107]. There has been some limited modeling of the relationship between the BSN and environment [119, 3, 112]. The main drawback of these models is that they typically focus on a specific issue in the BSN and are not easily extensible to consider other issues.

Despite these efforts, there is still a need for model-driven techniques for exploring issues like the effect of the location of the sensor on its output and the effect of particular user activities and environments on communication. Today, these issues are primarily explored using human subject experiments. Such methods are costly, especially for early concept exploration. In addition, there is a limited ability to keep some variables constant while changing others (*e.g.,* it is infeasible to have the subject reproduce the exact same

Conference on Embedded Networked Sensor Systems [7]. Some aspects of its design are also described in Scott Tepsuporn's senior thesis at the University of Virginia [101].

motion while a different sensor location is explored). Lastly, there have been recent calls for patient models compatible with device models to aid in the design of emerging medical systems like BSNs [2, 57].

The aim of the work presented in this chapter is to complement these experimentation techniques and respond to this call for patient models by providing a platform for carrying out experimentation and explorations in virtual space. The long-term vision is for a software platform which provides a researcher or designer access to a number of virtual human subjects on which he or she can place sensor nodes of varying capabilities and explore particular properties of interest in the system for various scenarios. This approach is line with the ideas developed in Chapter 4 on the need to explore variability when dealing with safety. It focuses on this idea for higher fidelity models.

The envisioned platform can be used in purely virtual fashion or as complement to human subject data. In the purely virtual case, there would be a repository of subjects and their behaviors that users of this platform can select from and run experiments on. In the complementary case, the user of this platform could collect the necessary information on a human subject's physical characteristics and behaviors, plug this information into the software platform to create a new virtual human subject and add virtual sensors to this scenario in order to carry out investigations.

This chapter presents the architecture and current instantiation of particular components for a multi-domain modeling and simulation framework (called BodySim) that embodies the basic features of the envisioned platform. It demonstration kinds of explorations that are possible with framework in the context of two motivating applications described below. BodySim leverages advances in 3D physical modeling and animation instantiated in open-source tools like Blender [16] and scientific computing tools like MATLAB/Simulink [103, 102] and Python (with appropriate libraries)[86]. We consider BodySim a multi-domain framework because the virtual human subject models describe a physical domain, whereas the models that provide information on the outcomes of interactions between the human subject and the sensors describe the interface between this physical domain and more computational and communication domains, as well as these computational and communication domains of the BSN. Figure 6.1 shows the BodySim concept.

It is important to note that a framework like BodySim does not seek to supplant previous modeling efforts for BSNs. Rather, it seeks to serve as framework where such models can interoperate and in particular interact with realistic models of human subjects. For example, the current instantiation of BodySim couples human models developed as part of this work

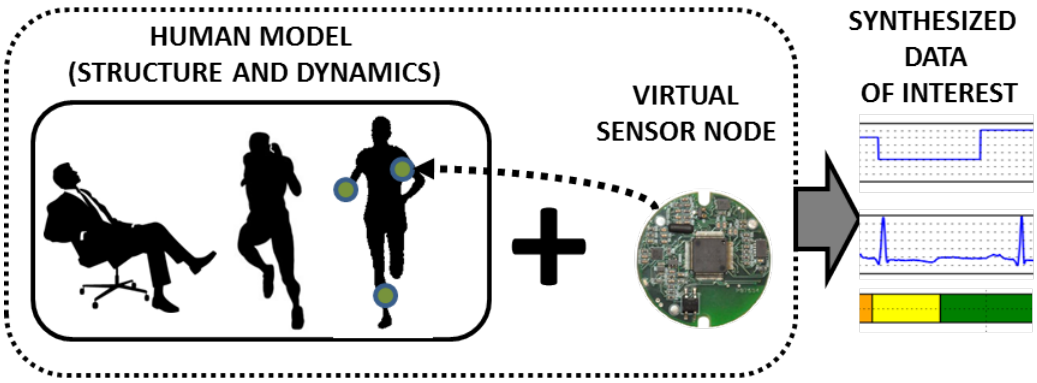



Figure 6.1: BodySim concept

with IMUSim [119], an inertial sensor modeling tool. The focus of this Chapter is not on the validity or the accuracy of the models like IMUSim used to explore BSN properties, but more to demonstrate the kinds of explorations that can been done when such models are coupled properly with virtual human subjects. The vision is that BodySim will be a community driven effort, and hence it is currently available as an open-source project online at http://wirelesshealth.virginia.edu/content/bodysim.

**Example 6.1:** Exercise Detection and Monitoring in the Artificial Pancreas

Exercise and other intense activities change the operating mode of human physiology in order to accommodate the increasing demands such activities place on the body. In a system like the Artificial Pancreas, this means that the way the body responds to insulin would change [19]. It has been shown that knowledge of exercise as capture by a body-worn sensor (in this case a heart-rate sensor) can be used to improve automated glucose management during exercise [20].

Inertial sensing is also another way to detect and monitor intensity of exercise, and the combination of inertial sensing and heart-rate can be used to detect variables related to intensity of activity like energy expenditure [27]. As seen in the meal-sensing example (4.13), misinformation the decision-making software can result in unacceptable risk. In chapter 5, we saw that physical factors can affect the information delivered to decision-making software by the BSN. Hence, it would be important to understand these effects and design the BSN to be robust to them as well as to pick

physical configurations that help the BSN reduce hazards related to its interaction with the decision-making software. We would like to know answers to questions like "which locations minimize potential for hazards?", and "which algorithms provide the robustness to a range of physical factors?" Such questions can be answered with an appropriate tool like the one envisioned earlier.

**Example 6.2:** Evaluation of Inertial Sensing Systems

Inertial sensing is an emerging sensing modality that is providing to be useful in a number of medical contexts [26, 80, 83, 25, 76]. As mentioned in the previous example, it can also be used to inform the Artificial Pancreas on when and how to adapt to exercise and other intense activities. Since sensing is heavily dependent on configuration, it is important to be able to evaluate the robustness of systems to various physical configurations as done by Gong *et al.* [40]. That work depended on human subject experimentation, and the envisioned tool could have enabled explorations of a wider range of situations with less overhead.

In addition, for some inertial systems, it is important to combine information from multiple locations to get a complete picture [26]. In this case, the effect of physical factors around the body on communication becomes important. The envisioned tool can allow this joint exploration of the effects of physical factors on communication and the effect of the resulting communication on overall performance of the system, to understand its robustness properties and susceptibility to creating hazards.

## 6.2 Basic Requirements

The aim for the envisioned platform is to be able to select a virtual human subject (and their environment), place sensors on them, simulate the human subject with sensors attached performing some activities, and obtain results on how the sensor behaves in the particular context the was simulated. A tool that enables this must have three main pieces. The first is human subject models which have information the spatio-temporal dynamics (how things evolve at each point) of the person with respect to the physical quantities important for the particular simulation. The second is an interface model that allows virtual sensors to be attached to the human subject model and which generates the physical inputs (and

others as necessary) to the sensor based on its configuration on the body and the dynamics of the body and its environment. The third is the sensor model, which takes as inputs information generated the interface model and generates on the behavior of the sensor given those inputs. These need for these pieces is implied by the generic BSN model developed in chapter 5

### 6.2.1 Human Subject Model ($H$)

The human subject model consist of two parts, the structure which describes the relevant three-dimensional setup of the body and the dynamics which describe how each point in that structure evolves in time with respect to the physical quantities of interest. For example, if the sensing scenario we are interested in collecting electrocardiographs (ECGs), then we would need the full body surface structure (mostly from the neck down) and the electrical properties at each point on that surface since electrodes are placed on the skin. We would not be interested in internal anatomical features. In addition, we would need dynamics that describe the persons motion and how different points on the body move. We would also need a model of how the heart functions for different activities, particularly, what the electrical potential at each point of the body given the movement of the body and the skin properties of the particular human subject. We may also include a model of electrical interference seen at each point. That way if the electrodes are placed on any point, we can extract the dynamics at that point and pass that to the sensor.

Formally, based on the generic BSN model in chapter 5, the human subject model ($H$) consists of a 2-tuple $(C_{S\to H}, Y_H = f_H(C_{S\to H}, t))$, where $C_{S\to H}$ describes the available points of interaction (the structure) and its properties, and $Y_H = f_H(C_{S\to H}, t)$ describes the dynamics at each point ($c_{S\to H} \in C_{S\to H}$) time at each possible point of interaction. Placing a sensor at a particular point there 'selects' the dynamics for that point to be used in the simulation.

### 6.2.2 Human-to-Sensor Interface Model ($H \leftrightarrow S$)

The human to sensor interface governs the interaction between the sensor and the human dynamics. It uses knowledge of the intended configuration to select and track the appropriate dynamics from the human subject model. In some cases it may process these dynamics to include additional physical effects. In our ECG example, this model would select the appropriate value of $c_{S\to H}$ for each electrode location and collect the dynamics generated

by $Y_H = f_H(c_{S\to H}, t)$. If we would like to incorporate motion artifact effects for example, the model could use knowledge of the movement and how well connected the electrodes are to modify the output from the human subject model to produce an electrical potential that accounts for these artifacts. If no additional physical effects are explored, this model then essentially passes the particular sensor configurations to the human subject model and passes the output from the human subject model directly to the sensor model. In this ECG case, the this model

Formally, the human-to-sensor interface model consists of a 3-tuple $(\mathscr{C}_{S\to H}, Y_{H\to S} = f_{H\to S}(\mathscr{C}_{S\to H}, Y_H, Y_{S\to H}), \{c_{S\to H}\} = f_{S\to H}(\mathscr{C}_{S\to H}, Y_{S\to H}))$. $\mathscr{C}_{S\to H}$ is the overall configuration information, which contains information on the specific points sensors are attached to the body ($\{c_{S\to H}\} = f_{S\to H}(\mathscr{C}_{S\to H}, Y_{S\to H}) \subset C_{S\to H}$) as well as other information necessary to generate inputs to the sensor model. $Y_{H\to S} = f_{H\to S}(\mathscr{C}_{S\to H}, Y_H, Y_{S\to H})$ describes the output to sensor model based on configuration information ($\mathscr{C}_{S\to H}$), the outputs from the human subject model ($Y_H$), and any outputs from the senor model back to the human subject model ($Y_{S\to H}$). Incorporating $Y_{S\to H}$ allows us to account for inter-sensor interference, especially in wireless communication. Remember that our main focus is on what happens to the sensor dynamics given the particular context and not the human subject.

### 6.2.3 Sensor Model ($S$)

The senor model describes how the sensor reacts to inputs from the human subject and its environment. These reactions could be in terms of internal properties like energy consumption, computation time, or outputs it produces that affect the human subject directly or for decision-making. In our ECG example, the model could contain a model of an analog front-end and analog-to-digital converter (including their noise characteristics) that converts information from the electrodes (through the interface model) to digital samples. It may even contain a signal processing model for filtering (in the analog or digital domain) the signal or doing some event detection. It could also contain a model of wireless communication or information storage behavior. In addition, it could include a model of the energy consumption of these components and the overall system energy consumption behavior.

Formally, the sensor model ($S$) is a 2-tuple $(\lambda_S, [Y_{S\to H}, X_S, Y_{S\to D}] = f_S(\lambda_S, Y_{H\to S}, X_S))$. $\lambda_S$ are the properties of the sensor that govern its intrinsic behaviors (*e.g.,* noise characteristics, processors speed, radio sensitivity). $Y_{S\to H}$ are the outputs the sensor produce the affect the human subject (and possible other sensors). $X_S$ is an n-dimensional state variable for various component properties of interest (*e.g.,* power or energy consumption, memory

used, actual data stored, transmission requests or packets received). $Y_{S\to D}$ is the information that the sensor generates for decision making. $f_S(\lambda_S, Y_{H\to S}, X_S))$ the behavior of the sensor in response to inputs based on its current state and intrinsic properties.

### 6.2.4 Overall Systems Model

The overall system model is shown in Figure 6.2. The user specifies the configuration, selects the specific human subject, and the specific sensor (also specifying the sensor properties). The various pieces then interact in feedback fashion in order to produce the right dynamics in the sensor model which can then be analyzed in conjunction with the trajectories of the other pieces that helped produced those dynamics.

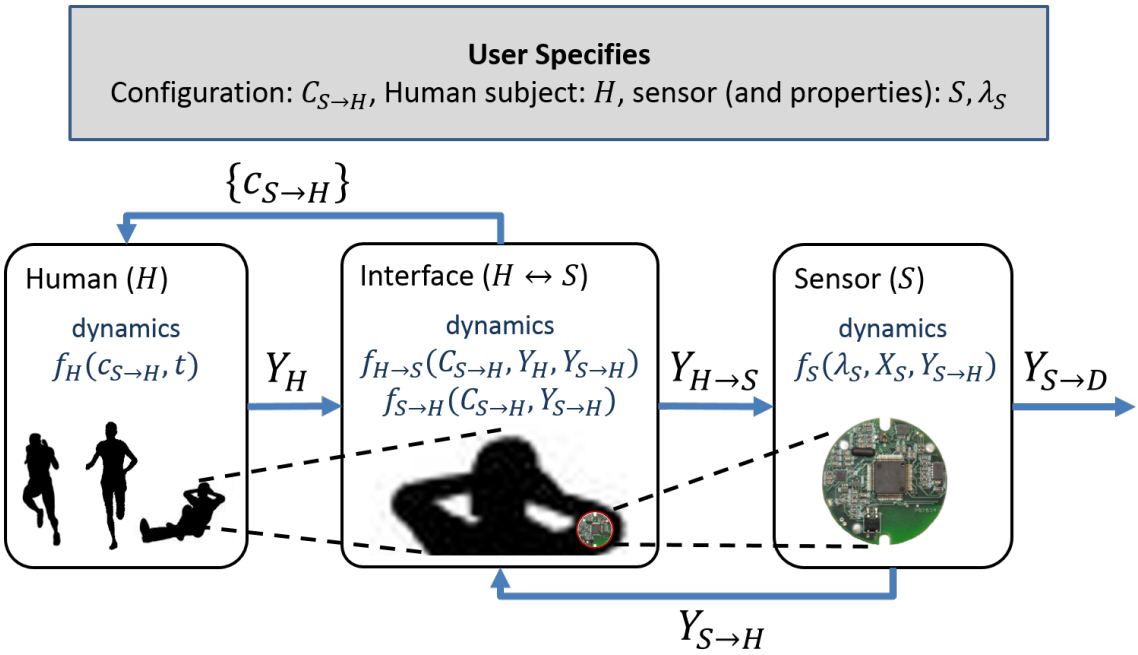



Figure 6.2: Overall systems model for design tools embodying features for exploring physical interaction effects on sensor dynamics

## 6.3 Realization in BodySim

The current instantiation of BodySim is focused on inertial sensing. It provides human subject models, interface models for inertial sensing and wireless communication, and ex-

tensible way of plugging in different inertial sensing and wireless communication models. It is based on the Blender 3D modeling and animation tool [16] extended with Python [86] scripts to provide the features described in the previous section.

### 6.3.1 Human Subject Model

The human subjects in BodySim are 3D object files (mesh files in the `.obj` format). The structure created from high-resolution laser scans (using a FARO Focus$^{3D}$ laser scanner [36]) of subjects. Figure 6.3 shows the setup for scanning subject and an example resulting 3D mesh. Each vertex on the mesh has a unique identifier which allows one to specify the location of the sensor.

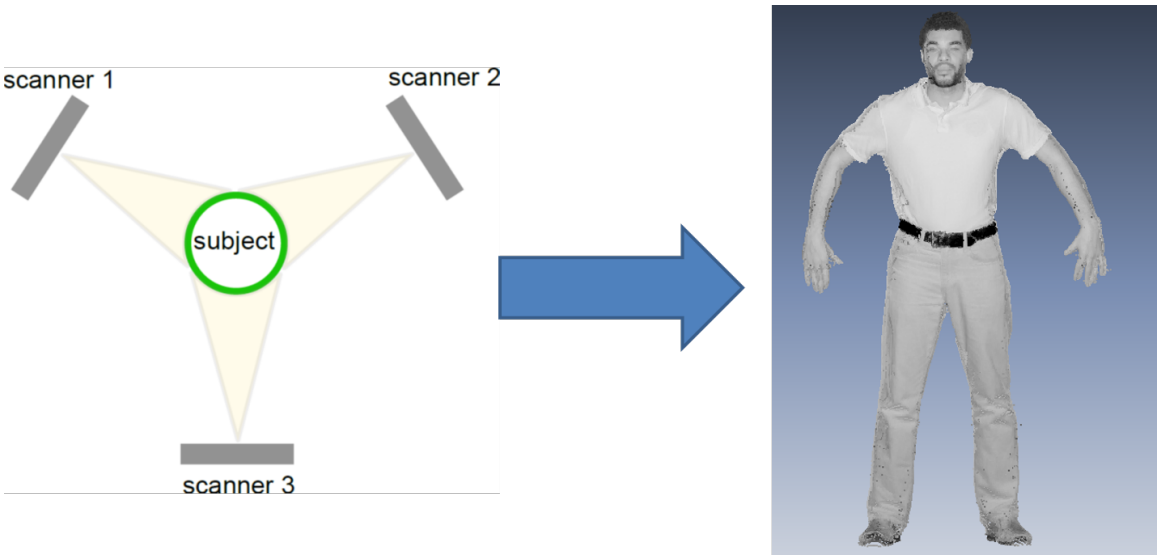


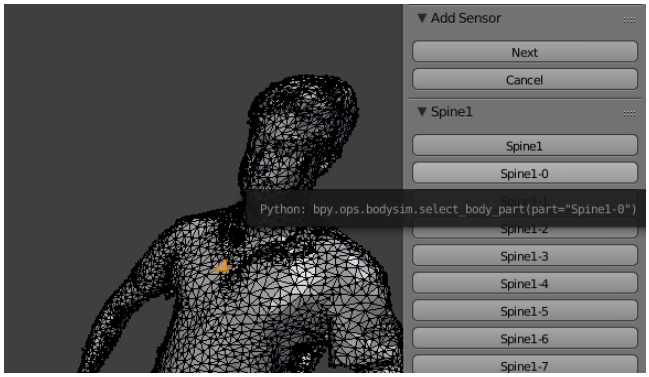

Figure 6.3: Setup for scanning human subject to obtain a structure model and the an example resulting 3D mesh.

The dynamics is represented by the same subject's motion captured by the Animazoo IGS180i motion capture suit by Synertial [100] (stored as biovision hierarchal format (`.bvh`) files). Each separate motion capture for a specific subject is a different file, though their scan is only collected once. All data on subjects (which is on-going) is being collected through an approved institutional review board (IRB) protocol. The dynamics of the subject is handled using the native animation engine provided by blender. The structure

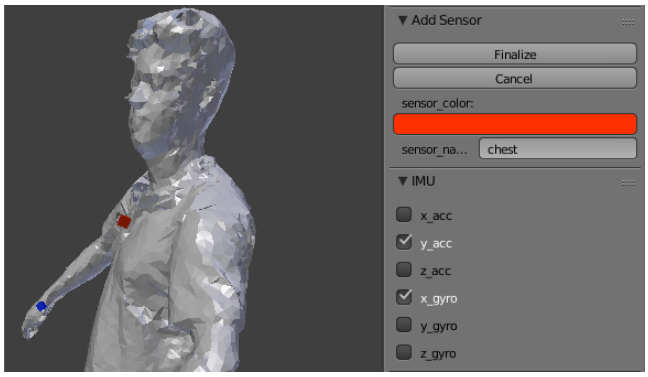
and dynamics of the subject are essential outsourced to the real world, to provide a more realistic model.

### 6.3.2 Human-to-Sensor Interface Models

Currently, there are two interfaces, one for inertial sensing and another for wireless communication. In both cases, the user can couple a sensor to the location on the body by selecting the particular vertex on the mesh and attaching the sensor there. The user can add multiple sensor to different locations this way. Using facilities provided by Blender, each sensor's motion is tracked based on the motion of the particular vertex it is attached to, which is controlled by the real-world motion captured for the subject.




(a) Choosing a sensor location.




(b) Choosing sensor parameters.

Figure 6.4: Adding a sensor to the human subject model.

In the current instantiation, sensor mounting is always correct and tight, so mounting issues like those raised by Gong *et al.* [40] cannot be explored. It straightforward however to provide an interface to the user for specifying the orientation of the sensor when it is attached to explore mounting orientation issues. Adding features for loose attachment would require some more work as the dynamics of this process needs to be understood. Lastly, the interface can produce as output the trajectories of the sensor that it tracked, which is useful for validating sensor models.

### Inertial Sensing

For inertial sensor, the main thing BodySim does is tracks the position and orientation trajectories of each sensor, making $Y_{H \to S}$ a 7-dimensional vector of these values (the orientations are tracked as quaternions [**?**]). These are generated during a simulation based on where the sensor is attached to the body and the movement of the particular subject. These are then passed to the inertial sensor model for its simulation. We track these variables because this is what the sensor model currently used requires. If another model requires the forces on the sensor due to the subject's motion, then we would have to provide an interface model that generates these forces from the human motion.

### Wireless Communication

The interface for wireless communication also tracks the position and orientation trajectories of sensors. In addition, it tracks two different parameters that are related to the body's effect on wireless communication.

The first is whether there is direct line-of-sight between any two sensors (it does this for all pairs of sensors). This corresponds to checking whether a line between two points goes through a vertex (a point on the body) or not. The second is parameter tracks how the body occludes (shadows) signals. It tracks how much of the transmitted signal escapes into the environment (and is not blocked by the body), which also analogous to how much of the signal reflected by the environment will be seen by the body. It is also based on checking whether lines between points go through the body. The basic idea (modified from what is described in Scott Tepsuporn's senior thesis [101]) is as follows.

1. Using the sensor's origin as the center, create a sphere whose radius is the height of the subjects body.

2. Create a number of 'equidistant' points on the surface of the sphere which correspond to discrete samples of the surface.

3. For each sample on the surface of the sphere, to check if there is line-of-sight between the point and the origin of the sensor (using the same method for checking if there is direct line of sight between two sensors but now treating the point on the sphere surfaces as the second sensor).

4. Find the number of samples that do no have line-of-sight to the origin of the sensor (because the body was blocking it).

5. Output the ratio of the number of non-line-of-sight points to the total number of points.

The above assumes that the antennas on the sensor are omni-directional and tries to find which how many of the possible directions are blocked the body. This parameter is tracked per sensor, and both variables are tracked for every time step in the simulation, creating a trajectory of each of these values.

The output to any wireless sensor model ($Y_{H\rightarrow S}$) depends on the number of sensors ($n$) in the simulation. Part of the output is the 7-dimensional vector of position and orientation variables plus the line-of-sight indicator. Another part is the $\binom{n}{2}$ pairs of line-of-sight indicator variables and the $n$ non-line-of sight ratios.

We track the two line-of-sight related parameters because the effects on wireless communication are different when there is direct line-of-sight versus when there isn't. When there is direct line of sight, the effects on wireless communication is dominated by the relative position and orientation of the sensors with some contribution by reflections of signals by the environment. When there is not direct line of sight, the communication is dominated by reflections of signals by the environment.

### 6.3.3 Sensor Models

The only functional sensor model in the BodySim currently is the inertial sensor model. To demonstrate the extensibility of BodySim and its interoperability with available models, we used an open-source model called IMUSim [119] developed by Young *et al.* , who showed its ability to produce accurate simulation results for the inertial sensor they developed [120].

IMUSim is an open-source model developed in Python, which models a 9-degree-of-freedom inertial measurement unit (accelerometer, gyroscope, and magnetometer). It provides both an idealized model and the ability to configure it to mimic realistic systems. The model provides a way to configure analog-to-digital converter parameters, timing issues in sensors, and even prototype signal processing algorithms, and communication of data between sensors. For simulations, it needs as input the trajectory (position and orientation) of the sensor, which BodySim provides. Figure 6.5 shows an example output for a simulation using the ideal inertial measurement unit model.

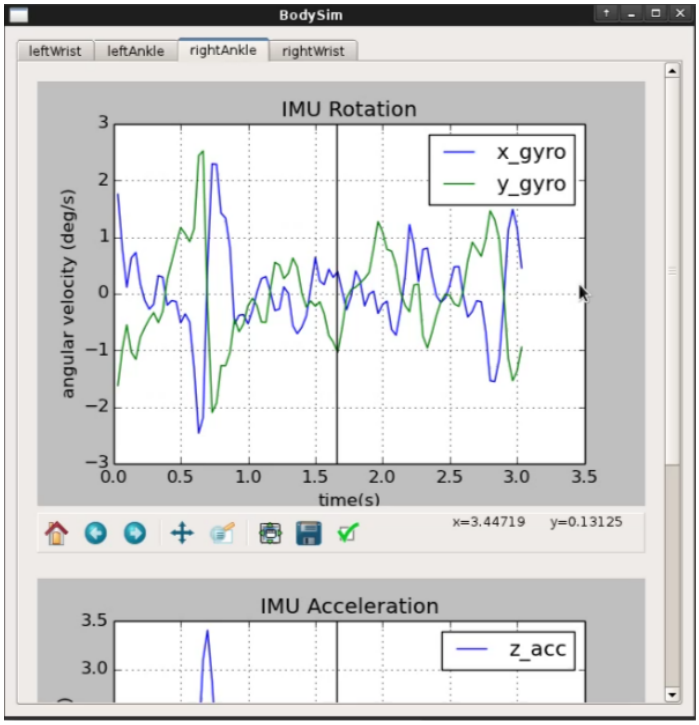



Figure 6.5: Example output for ideal inertial sensing simulation.

BodySim provides the ability to select which

### 6.3.4 Overall Simulation Flow

In BodySim, the human subject models are provided. A user selects the human subject (which would also select the particular motion that was collected for the subject). The user can then configure the setup by selecting the sensor location and the properties of

the sensor the user wants to observe, which in this case corresponds to which axes of the accelerometer or gyroscope they would like to observe. One this is all set up, the user can run the simulation which would first generate the sensor trajectories and pass these to IMUSim for simulation and record all outputs into human-readable (`.csv`) file format.

### 6.3.5 User Interface Features

BodySim provides a number of features to make working with the models easier. The overall user interface with a few features is shown in Figure 6.6.

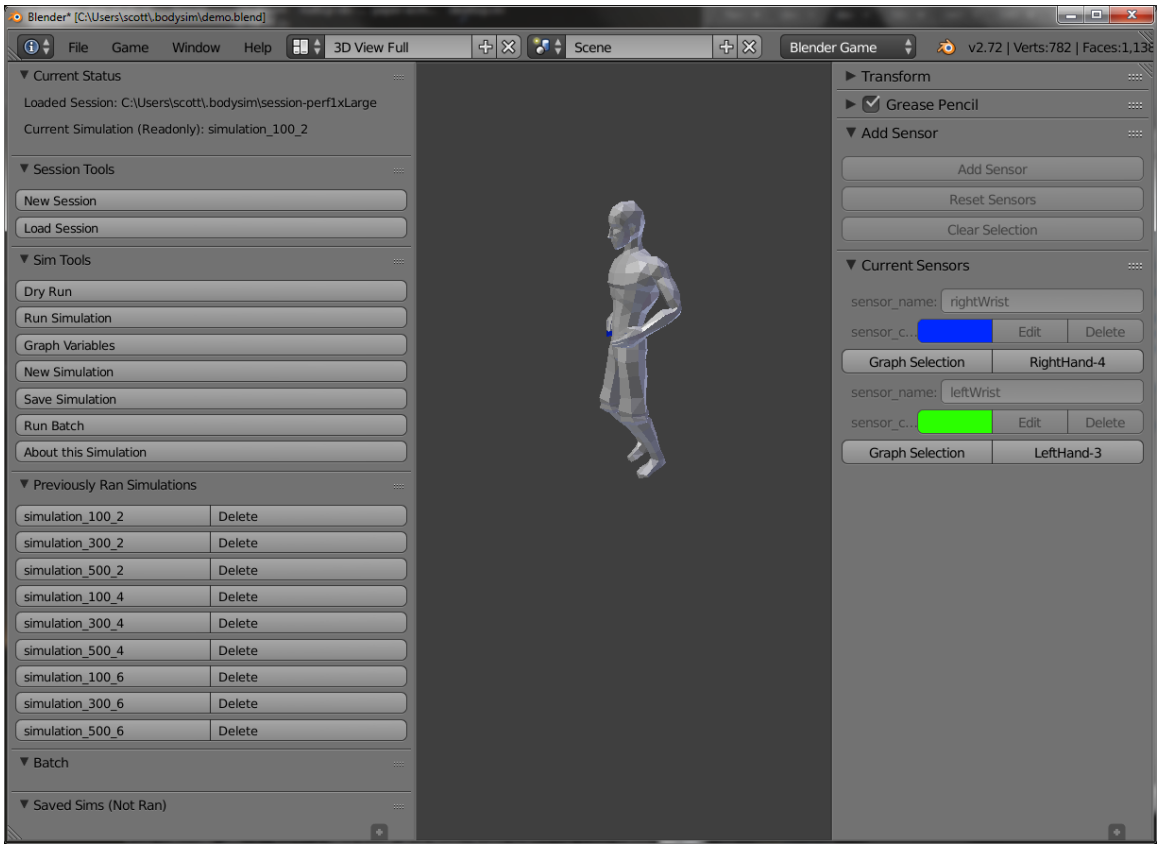



Figure 6.6: The BodySim user interface

#### Human Subject Visualization

The first important feature is the ability to view the human subject, see where sensors are placed, and watch the experiment if desired. This visual interface provides an intuitive feel and gives the user the sense of running virtual experiments just like they would a real experiment. In addition, before running a simulation, the user can do a 'dry run' to see what the motions for the particular subject are.

#### Sessions and Simulation Configurations

A number of related simulations can be grouped into a session. A simulation consists of the the particular human subject, and the sensor configuration. If it has already been run, then it would also contain the data from the simulation run. A new simulation can be created by copying the configuration of a previous simulation and using that as a starting point. There is also a facility for saving simulation configuration and running them later, or adding them to a batch of simulations to be ran as a group later.

#### Sensor Configuration

In addition to adding sensors to a specific location and selecting which axes to simulate, user can also name sensors and color code sensors so they can identify them visually on the model. The sensor names are also used as labels for the data which helps with post processing. If the user does not provide a name, the sensor is given the name of the location of the body it is attached to.

#### Graphing of Data

Once a simulation has been run, users can graph the data that was produced. In addition, the graphing interface is setup so that when the user clicks on the timeline, it shows what the subject was doing at the time that particular data point was produced.

## 6.4 Utility of BodySim

BodySim has a number of interesting applications. We are currently exploring one, and other potential ones are detailed below.

### 6.4.1 Wireless Communication Model Development

It turns out that BodySim is crucial to the development of wireless communication models. We already mentioned that understanding how the body occludes sensors is important. This occlusion is near-impossible to track in an actual experiment, even if video ground truth is recorded. It certainly cannot be done purely from motion capture data correlated with wireless communication data, since motion capture systems only provide a skeletal model of the person. However, like we have done, BodySim, we can track this occlusion in a realistic 3D model of the subject automatically, and use these as parameters to fit the wireless data to for the wireless communication model, which is what we are doing in on-going work.

### 6.4.2 Virtual Prototyping

As mentioned in chapter 4, one of the big issues with health management systems in inter-person variability. Hence in systems designs, it is important to test ideas across a number of different people representative of the intended use population. BodySim provides the ability to do that in a low overhead way with virtual models of subjects and prototypes. Because it provide multi-domain information, a full design can be simulated (including algorithmic and communication considerations). Once promising candidate prototypes are identified, then those can be tested on real subjects.

This prototyping and experimentation need not be only with models. A hardware-in-the-loop simulation could be run with BodySim providing the physical environment information to the hardware and receiving information about physical actions of the hardware.

### 6.4.3 Simulation-Based Design Space Exploration

The motivation applications presented earlier demonstrate some of the design space exploration needs for BSNs. In both the exercise-informed glycemia management case and the inertial sensing case, a number of different configurations need to be explored in order to understand which configurations and algorithms go best together. In particular the designer would be looking for algorithms that are most robust to variability and configurations and physical interaction issues, as well as reasonable sensor locations. BodySim can provide the data needed to explore the performance of the system for various configurations and situations in order to develop concepts for prototypes which could then be tested using more elaborate systems models in a virtual prototyping exploration.

### 6.4.4 Benchmarking

From a regulatory perspective, BodySim could be a good benchmarking tool. Many inertial sensor experiments used to increase confidence in design use different subjects. Hence it is difficult to compare the results of one groups sensor and experiment to another's since both were not run on the same set of subjects. BodySim could provide a common cohort of subjects for designers and manufacturers to benchmark against. In addition, the FDA could work with independent groups to provide a repository of human subjects that are representative of the expect use populations, on which designers can test new concepts before going further in the design process.

## 6.5 Relation to Other Approaches

Many modeling and simulation tools available either are too general (not specifically targeted at the BSN domain) or focus only on a specific aspect of BSN design. In the first category are tools like MATLAB and Simulink from the Math Works, Inc. [103, 102] and Ptolemy from the University of California, Berkeley [35]. One could conceivably perform a number of the explorations described above, but modeling especially human subjects would be cumbersome.

In the second categories are tools like ns3 [1], OpNet [90], and castilia [72], which focus mainly on networking and communication, and TOSSIM [62, 56, 33] and IMUSim [119], which mainly focus on sensor modeling. These tools mainly lack the appropriate environment model provided by something like the human subject model and the interface model in BodySim in order to provide realistic about how interactions with the body actually affect sensor dynamics

The main advantage of BodySim over these tools is its more realistic environmental model made up of the human subject models and the interface model. Apart from that, the sensor models could be prototyped in these other tools and the BodySim could interoperate with them to provide the appropriate outputs. In some cases, parts of the interface model could be prototyped in other tools like MATLAB or Ptolemy. Modeling the effect of loose mounting in inertial sensors is an example of such a case. As mentioned previously, the aim of a tool like BodySim is not to supplant previous approaches, but to use insights from the generic BSN to provide the missing pieces and a framework the allows these various efforts to be integrated.

## Summary

Herbert Simon in his book, "the Sciences of the Artificial" [95] discusses the role of the operational environment in shaping design. He talks about sundials performing as clocks in sunny climates, and clocks used on ships having much different designs than those used in the home. This resonates with ideas raised by the generic BSN model about the need to explore the effect of the interaction of the BSN with its environment: no matter how good the models of our design are, without a good environmental model to explore its actual performance, we risk ending up with a poor design. The developments in this chapter focused on this issue of ensuring that tools enable this crucial aspect of the design for BSNs, especially because of the dynamic nature of the environment in which BSNs operate. The BodySim software framework described provides a proof-of-concept on how to make these features available. Not only does it include a realistic human subject model, it, more importantly, ensures through the interface model that this model can interoperate effectively with models of BSN designs: good models of the environment and good models of the design are useless if both cannot interact.

# Part III

# Epilogue

# Taking Stock

> Follow effective action with quiet reflection. From the quiet reflection will come even more effective action.
>
> Peter Drucker

It is important to pause and reflect on the ideas presented in the previous chapters, highlight the key points, and draw some conclusions.

The overall aim of this dissertation was stated as to provide mechanisms for thinking about what it means for an emerging computer-based medical technology to be safe for patients. Given a technology, or concept for one, we wanted a way to examine the potential for harm in a systematic manner and come to some conclusions on whether this is acceptable or not. We also wanted to be able to use the insights from this method of examining technologies to inform design so we can ensure that we end up developing technologies that are acceptable.

## Core Ideas

One contention was that even though the aim systems safety is to deal with this issue of potential for harm for various systems, its techniques were ill-equipped to deal with emerging medical technologies because the technologies they were primarily developed for had different goals than medical technologies. Much of the dissertation was focused on using the principles of systems safety (not the techniques) to develop mechanisms for reasoning about patient safety of medical technologies that we consistent with system safety ideas (hence, complementing existing techniques) and that were valid for the health context. Below is a recap of this development, highlighting the key insights and the main takeaway points.

## Systems Safety as Reasoning about Emergent Behavior of Dynamical Systems

We began with the need for our framework for reasoning about patient safety that is both valid from a health management perspective and consistent with systems safety principles. The approach was to put the systems safety goal of achieving "acceptable mishap risk" in health terms, but in way that where these could also be linked to systems design issues. The key was in the dynamical systems view of both health management and systems safety. Chapter 2 established that this view is present implicitly or explicitly in many systems safety paradigms, and that the main differences were in how each paradigm viewed the mechanism for emergent behavior.

It showed that mishaps were related to events, which could be considered values that were part of the trajectories of dynamical systems, that risk was related to the possibility of particular events (the undesirable ones) occurring and hence the system exhibiting particular undesirable trajectories. In addition, risks were related to the severity of these undesirable events. The main differences in paradigms were in how they reasoned about the causes of increase in possibility of undesirable events. The linear approaches focused on particular components (usually humans or mechanical components), assuming that these were the dominant drivers of the emergent behavior. The non-linear approaches took a more holistic approach, focusing on the relationships between interactions and the emergent behavior that produced undesirable trajectories.

The key takeaway was that regardless of the particular system safety paradigm, the main concern was with understand what the undesirable trajectories (mishaps) are, how they might arise (hazards), and intervening through design to reduce their possibility and severity (risks) to acceptable levels. This is summarized in figure 6.7

## The Human Body as a Natural Safety-Critical System

Chapter 3 showed that the dynamical systems view was a valid one for human function. With this link in mind, chapter 3 connected human function to systems safety further by viewing the human body as a natural safety-critical system, with health being the body maintaining safe states or exhibiting desirable dynamics (acceptable function) and body failing to exhibit these dynamics (unacceptable function) as the reasons for intervention. Mishaps were therefore these transitions to unacceptable function. This translated the dy-

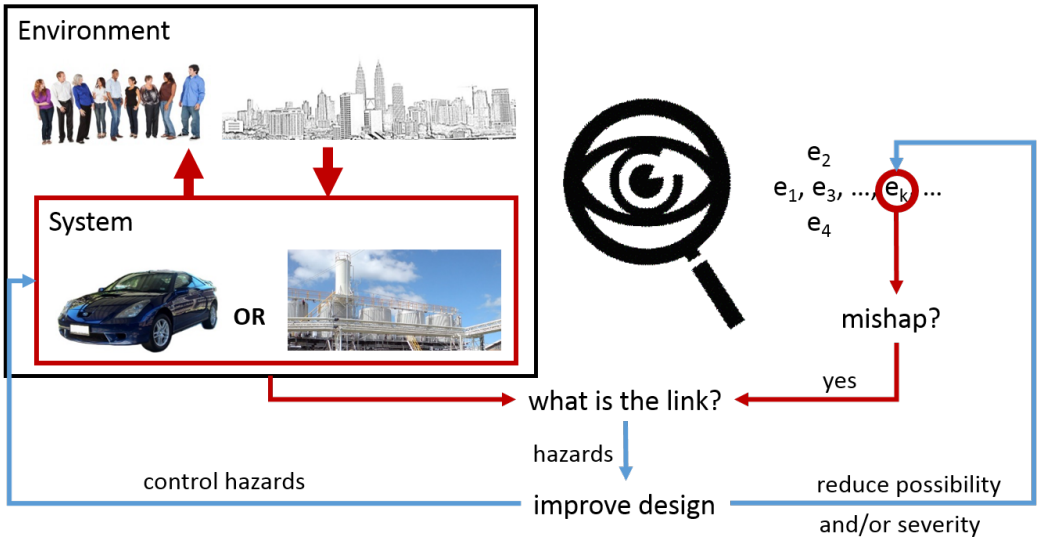



Figure 6.7: Illustration of systems safety principles.

namical systems view of safety critical systems to a view of human function as shown in figure 6.8.

### Health Metrics as Risk Measures for Human Function

Chapter 2 had introduced the idea of risk the possibility and severity of events that led to loss. In particular, we saw that if we assumed that we could identify these mishap events ($e_k$ such that $f_{\text{mishap}}(e_k) = \text{mishap}$), then we could have a risk function for these events, $\text{risk} = f_{\text{risk}}(p(e_k), q(e_k))$, where $p(e_k)$ is the likelihood of that event occurring, and $q(e_k)$ is the severity of the mishap.

Chapter 3 showed that health metrics played the role of these risk functions. In particular, they were typically interpreted as indicators of what likelihood was of transitioning (immediately or in the long term) into states with more severe consequences, based on information on the current and/or past trajectories of the human dynamical system. This development is summarized in figure 6.9

### Health Management as Safety Interventions

Chapter 3 made one more link to systems safety ideas. Once we are able to measure risk, the goal is to control it; to reduce it to acceptable levels. This is where medical technologies

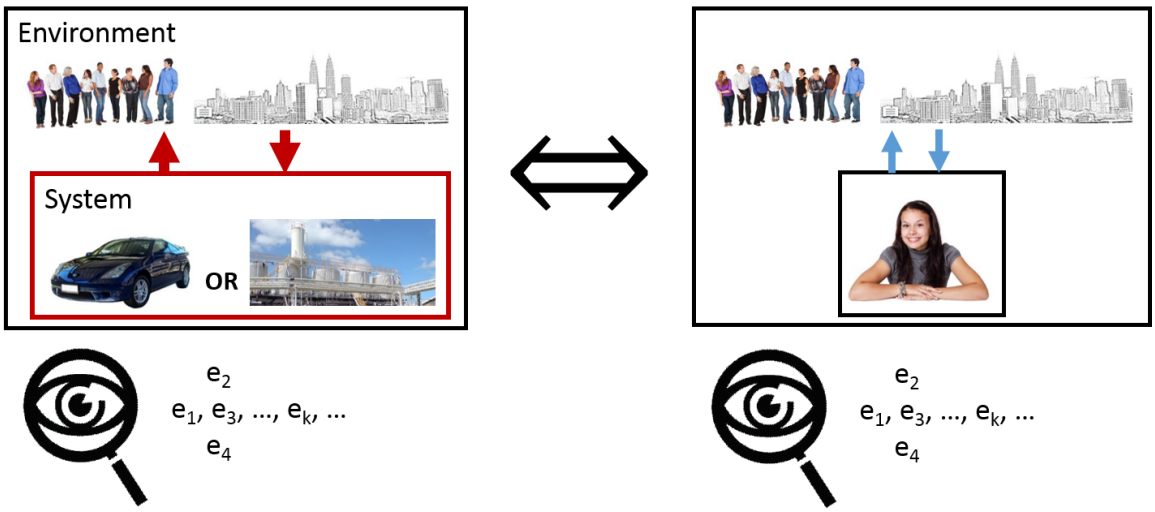



Figure 6.8: The link between the dynamical systems view of safety-critical systems and that of human function.

come in. They form part of a feedback system designed to influence the dynamics of the patient so that risks (potential for undesirable health outcomes) are minimized. This link between controlling hazards in a safety-critical system and health management is illustrated in figure 6.10.

The key takeaways here were that the human body is a natural safety-critical system which the potential to enter unsafe states. The way we evaluate this risk is through the use of health metrics. When the risk is high or when the person has transitioned in unsafe states, we intervene with the aid of medical technologies to reduce the risk or influence their dynamics to evolve back to safer states.

## Patient Safety of Medical Technologies as Robustness to Variability

With the notion of health metrics as risk measures, and health management as risk control, chapter 4 considered mishaps in the case where medical technologies are introduced. It showed that adding medical technologies creates new dynamics with the intention of moving the patient towards safer states, but there was the potential to also introduce unacceptable dynamics. In particular, it focused on the need for medical technologies to be robust to the inevitable variability exhibited by the people they are intended to help, which

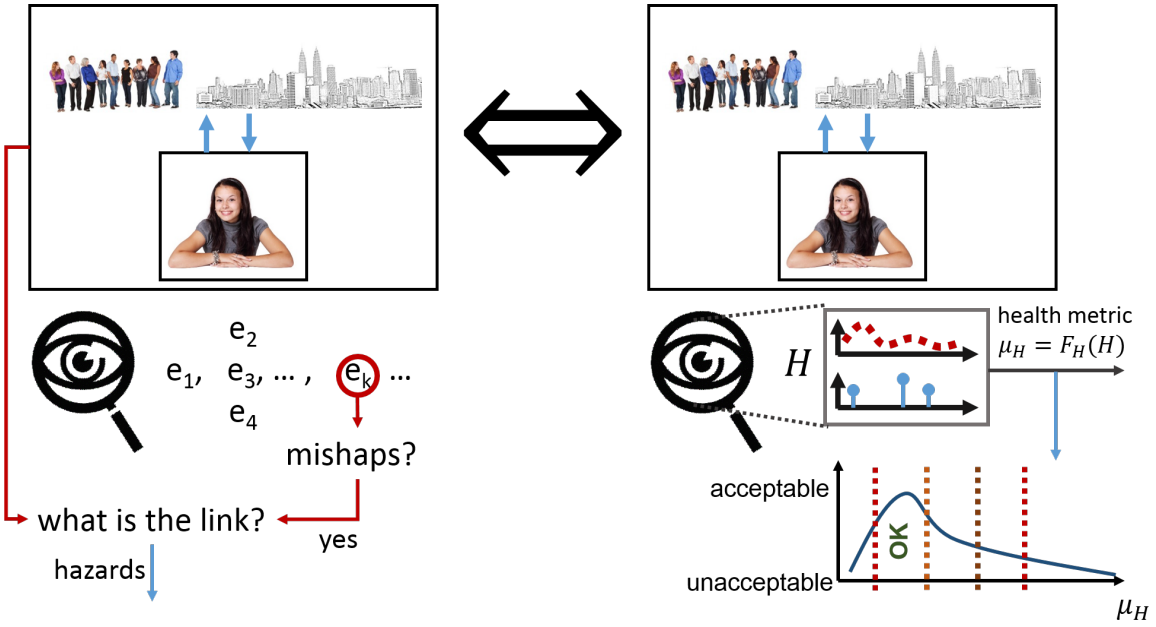



Figure 6.9: Illustration of health metrics as risk measures.

was what introduced newer risks (the possibility that in use by a patient mishaps could occur due to the introduction of the technology).

Acceptable mishap risk in this context was then defined as a function of health metric outcomes across a representative population relative to some baseline. The use of a population was the way of accounting for variability in patients. The baseline essentially represented a reference risk level outside the particular health management context where the technology under consideration was used. We wanted outcomes overall to be better than the baseline, and the variability in outcomes (especially in the direction worse than the baseline) to be minimum.

What we essentially did was define patient safety of the technology as a relative risk accounting for variability in risk outcomes. We maintained the health metrics as our risk measure for individual patients, allowing us to keep patient outcomes central to the model of patient safety. The baseline represented a reference risk level (in terms of health metric outcomes) when the technology under consideration was not in use. When a set of baseline values was used, this accounted for variability in risk in the baseline case. The health metric outcomes from 'testing' the technology on a representative population represented the new

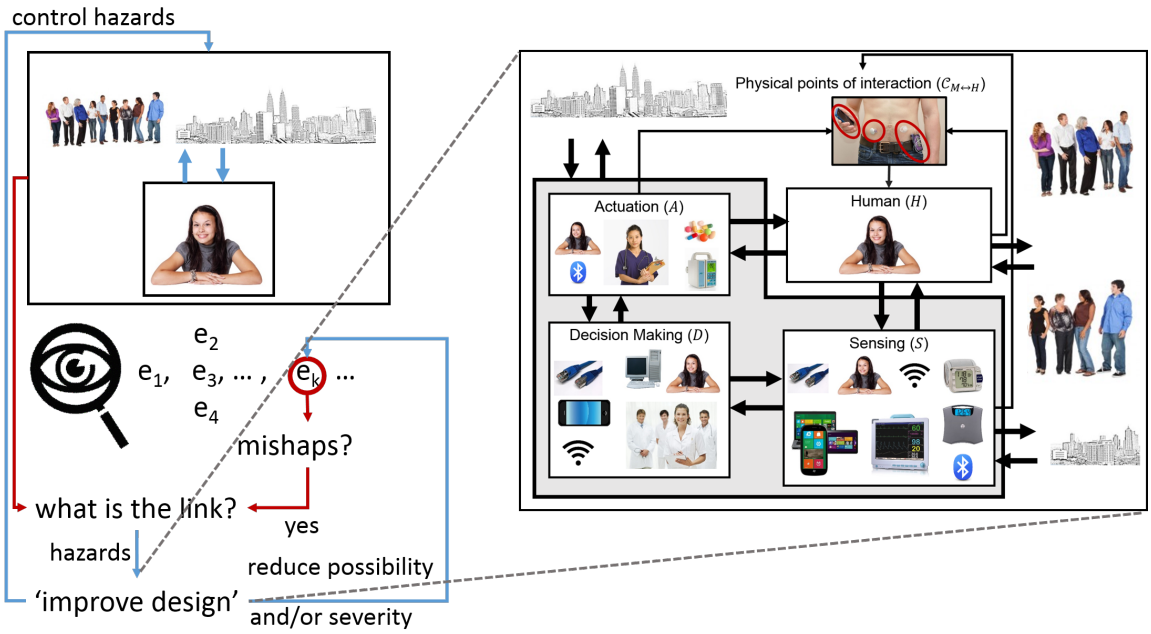



Figure 6.10: Illustration of health metrics as risk measures.

risk level as a result of introducing the technology accounting for the inevitable variability as well as the potential for the technology to introduce new risks.

Determining whether this risk was acceptable or not was done through comparing the new risk level to the baseline risk level (using the population-level risk metric). The population-level risk metric essentially measured how well the technology was able to help in controlling the original (health) risk it was designed for, and how well this was balanced with introducing newer (health) risks. The acceptable risk criteria determined the boundary between acceptable and unacceptable in terms of this population risk metric.

## Utility to Stakeholders

The main utility to stakeholders of the model of patient safety is in making precise and explicit the pieces that go into reasoning of patient safety of a particular technology. The idea of the patient safety criteria detailed these pieces. The health metrics of concern and their interpretation must be identified. These determine the hazards to be 'controlled' by the health management context in which the technology is to be used. A baseline must be established to compare outcomes for the new technology to. Variability must be accounted

for in all pieces (especially the patient population) and it must be stated how this is done. How the baseline and the new risk level are to be compared must be established.

With these pieces explicitly stated, stakeholders can have meaningful discussions about the criteria to update or modify it. This is important in allowing all stakeholders to reach a common understanding of safety for a particular technology. In addition, with the right information, a technology can be assessed for patient safety at various stages of the design by a designer or regulator; the parameters affecting the dynamics of the technology can be linked to risks and used in safety-guided design; different subsystems of the same health management context can be considered making assumptions about the behavior and variability of other subsystems; and the sensitivity of a specific design to assumptions made in any of the criteria pieces can be explored.

# Implications and Applications

The core ideas were developed in a very general context, focusing more of defining patient safety than design issues. Below is a recap of the developments that focused more on design issues in the context of body sensor networks, an emerging technology that embody the complexity-increasing trends of autonomy, integration, and mobility.

## Patient Safety Analysis of Body Sensor Networks

One of the keys to reducing patient safety issues is the identification and controlling of hazards. Hazards are those design features that are related to the risks. In the case of medical technologies, we need to link the outcomes to design parameters.

Focusing on the functional abstraction approach of non-linear paradigms systems safety paradigms, chapter 4 indicated that a way to link outcomes to design parameters that it was first important to understand the black-box behaviors. This allowed for reasoning about the properties of different possible implementations. With the black-box behaviors for the subsystem of interest, one could then explore the relationship between variability of the parameters and variability in the population level risk in order to determine the acceptable behaviors.

Chapter 5 provided a black-box model for BSNs, and used this model to develop a set of generic behaviors related to the intrinsic behaviors of the BSN. Using these behaviors, it developed a generic set of hazards at the human-to-BSN and BSN-to-decision-making interfaces, where were the interfaces where the BSN could affect the health management

system in ways that produced undesirable outcomes. It then showed that the black-box parameters of interest for safety-guided design for BSNs were those related to the hazards. In essence, in order to derive the hazard criteria, one had to explore the relationship between the variability in the hazard-related parameters and the population risk as shown in chapter 4, and use the acceptable risk criteria to determine the acceptable behaviors (which would then be the hazard criteria). One could then examine the parameters of different design choices and see if they resulted in hazards or not. In the spirit of systems safety, there was a discussion of potential causal factors for the hazards and some points to keep in mind in a safety-guided design of BSNs.

### Desing Tools for Body Sensor Networks

One of the main points that came out of the patient safety model development in chapter 4 was the importance of variability. From the perspective of the BSN, the patient and their environment is the operational environment, and this is the significant source of variability that the BSN must account for. In particular, physical interactions are especially important to consider, since the BSN relies on these interactions to achieve its function, but these interactions can also interfere with the intended function. The tools the BSN designers use must therefore enable exploring these issues.

Chapter 6 showed a proof-of-concept tool that embodied this features for inertial BSNs. In particular, it used realistic 3D models of humans and their motion captured using 3D scanners and motion capture equipment as the human model. In addition, it provided interfaces for connecting these human models with virtual inertial sensing models to enable the explorations necessary for understanding the effect on physical issues on sensor dynamics and potentially on hazards. It's main advantage is in providing more appropriate environmental models and an interface to allow these model to interoperate with sensor models. In addition, it allows multiple issues (like processing, communication, and energy consumption) with sensors to be considered in the same software framework, providing more insight on overall sensor dynamics than tools that focus on individual issues.

## Connecting the Different Pieces

Some links were made between the different pieces, but the focus of the dissertation was not how they all fit together. The brief discussion below puts the pieces together in the context of BSNs for the example in exercise monitoring for the artificial pancreas.

**Example 6.3:** Exercise Detection and Monitoring in the Artificial Pancreas

Recall that exercise and other intense activities change the operating mode of human physiology in order to accommodate the increasing demands such activities place on the body. Breton *et al.* showed that knowledge of exercise as capture by a body-worn sensor (in this case a heart-rate sensor) can be used to improve automated glucose management during exercise [20]. In addition, Chen *et al.* showed that the combination of inertial sensing and heart-rate can be used to detect variables related to intensity of activity like energy expenditure [27]. Let us assume that we would like to build a BSN that senses motion to provide exercise information to the decision-making software in the artificial pancreas.

**Patient Safety Criteria** The first thing we need according to the patient safety model in chapter 4 is the patient safety criteria. Central to this is the health metrics. In terms of functional and side-effect outcomes, we are interested in blood glucose levels. We would have metrics associated with blood glucose levels. If we had information on the choice of materials of the BSN components as well, we would have metrics associated with physical interactions like potential for skin irritation.

Our system setup is made of up the BSN, the other pieces of the artificial pancreas (the decision-making software and the pump), and the patient. We would need some idea of how these pieces behave (black-box behaviors would do). We would determine what our baseline population would be. In this case it could be information on health metrics from artificial pancreas users without the exercise information. We would come up with a population risk metric that would compare the outcomes from our system to the outcomes from this population, and an acceptable risk criteria.

**Safety Requirements for BSN** With the criteria, we could characterize the BSN in a black-box manner using mechanisms provided in chapter 5. With coarse models of how behaviors of the system result in outcomes, we could explore the space of these parameters and determine the acceptable behaviors using the acceptable risk criteria to identify those parameter values that result in acceptable risk. This exploration could also a combination of models and data from pilot trials if feasible and cost-effective. If the criteria needs to be relaxed to come up with a reasonable acceptable behavior region for our requirements, we would have this discussion with health practitioners

and/or the regulators, or we could make a choice and document our rationale. With these requirements, we would have the hazard criteria for our BSN components.

**System Design Explorations** With the hazard criteria, we would then like to develop design that meet the criteria. A tool like BodySim in chapter 6 would allow us to test details of different virtual prototypes across a population of subjects to understand how robust our algorithms are to physical and communication issues. We may end up with an algorithm that is agnostic to where the sensor is placed, or the depends on a particular location but presents information to the decision-making software in a way that avoids hazards. We would identify promising designs by examining the sensor dynamics related to the hazard parameters and computing what parameter values they result in. If the values satisfy the criteria that the particular design point is acceptable, and if not then we would reconsider the design. Once we have what we think is an acceptable point, we would like to test it in a higher fidelity scenario to increase our confidence in the design.

**Assessment** Once we have an acceptable system, or if we were given a candidate system to examine, we gather information (through a real trial on an appropriate population) on how the system does with respect to the particular metrics in the assumed, or intended environments. Once we have the information on the metrics, we would compare them to the baseline metrics according the the population risk metric and determine if its risk falls within the acceptable risk region. If it does, we would consider the system acceptable for use. If not, we would understand how it failed to fall within this region and use that information to help with a redesign. We could also have a discussion on whether the criteria should be relaxed if the system is close enough to the acceptable region.

From the above, we can see that the safety criteria is central to the design and evaluation of the system. A key part of the criteria is the health metrics. First is the functional metric, which indicates the 'natural hazard' the medical device must help control. The second is the side effect metric, which acknowledges the possibility of the technology to introduce newer hazards. Chapter 3 provides the mechanisms for specifying and discussing health metrics. Chapter 4 provides the mechanisms for precisely and explicitly specifying the criteria based on the health metrics. Even before a design, or during an assessment of one, the criteria is available and can be discussed.

The link between acceptable behaviors and the criteria made in chapter 4 then helps a designer to develop requirements. For a system like the BSN, ideas in 5 help focus on the important parameters. This allows designers and other stakeholders to refine the requirements and the criteria, if necessary, to balance feasibility with potential for and severity of mishaps. Once the requirements are developed, candidate designs can then be explored. Since requirements are linked to the criteria, if the this process reveals the need to alter requirements, then another discussion can be had about the requirements and how this affects the criteria. A tool like BodySim presented in chapter 6 then allows higher fidelity explorations of the behavior to refine and flesh out designs. If new issues arise, we can always use the links to go back up the chain all the way to the criteria to see if needs to be adjusted.

# Potential Future Directions

> Every new beginning comes from some other beginning's end.
>
> Seneca

The developments presented in this dissertation are certainly not the final word on any of the topics. Many of them present starting points for interesting future work. Below are some thoughts on possible future directions.

## Health Management Model

The structure and interpretation of health metrics presented in chapter 3 provide an interesting framework for examining existing health metrics, the rationale behind them, and how their interpretation factors in intervention decisions. It is provides a scheme for developing health metrics for emerging technologies. It would be interesting to explore this scheme for a number of diverse health metrics.

In particular, the idea of proxy metrics that are readily observable during health management and their relationship to other metrics that would be used to measure actual health management outcomes but are not feasible to measure frequently. For example, in cancer treatment, many of the 'metrics' for tracking progress require imaging, blood tests, or biopsy analysis, each of which can be costly and in some cases uncomfortable for the patient to obtain frequently. A proxy metric, could be another variable that is easier to obtain more frequently but provides enough information to help track effectiveness of treatment.

Also, the idea of the body as a natural safety-critical system seems an interesting one to explore as a perspective for thinking about health and health management. In particular, understanding its repair and robustness mechanisms, could help develop technologies that work synergistically with these existing mechanisms, and reduce the potential for issue that happen when there is interference between these and the medical technology.

## Patient Safety Model

One obvious next step is to use the framework to guide the end-to-end design of an emerging technology and organize its safety argument. This was beyond the scope of the dissertation. One such emerging technology where there are the necessary resources to do so is the artificial pancreas. There are other areas like cardiovascular disease management. In general, the areas that would benefit most are those for chronic disease management where the patient is in consistent interaction with the health management system.

Another step is to work with stakeholders to see how this framework can be incorporated in guidance and policy provided by the FDA and how medical technologists can fit the ideas into their specific work flows. The end-to-end case study would provide some insights into how this might work. The FDA sometimes organizes mock submissions to help test new approaches, and a mock submission of an emerging technology would help. This is a process that could take quite some time, but is worth exploring.

## Body Sensor Network Patient Safety Analysis

Similar to the above, an obvious next step is to work with a designer on an incorporating these ideas into their workflow. In particular, it would be interesting to explore how the hazard definitions can be used with formal verification techniques and other model-based design approach to identify issues with designs before committing to particular implementation details.

## BodySim

On-going work with BodySim is to provide a wireless communication model. Like, the above, BodySim could be used in virtual prototyping explorations of different body sensor networks including inertial sensors. The wireless communication model could actually be used with body area networks as well, where part of the human model exists outside of BodySim to account for effects of actuation. This interoperability with other models would certainly be a key issue to explore to harness the advantages of BodySim. One such integration is with the OpenSim [32] gait simulator target at simulating abnormal gait. These gait models could be couple with virtual subjects in BodySim to prototype system

for monitoring abnormal gait to help with identifying diseases that manifest as abnormal gait or to help health management where improved gait is used as a proxy health metric.

The current efforts for BodySim focused more on providing the features and infrastructure. There is quite some work to be done on the user interface, and performance of simulations. In addition, there is the question of whether BodySim should depend on Blender for the 3D portions, or should use another, possibly more stable, but still accessible, platform to support the physical model aspects. With the write collaboration, the physical modeling could be written from scratch.

## General Systems Safety

One interesting direction in the general area of systems safety is consider this idea that at the core, most system safety techniques are dynamical-systems-based, but differ in their views of emergent behavior. A review of paradigms, techniques, and the history of systems safety using this framework would be useful to put various techniques in context and help practitioners sort through the suitability of different techniques for various design tasks.

# References


[1] ns3. [Online]: https://www.nsnam.org.

[2] High-confidence medical devices: Cyber-physical systems for 21st century health care. Technical report, The Networking and Information Technology Research and Development (NITRD) Program, February 2009.

[3] T. Aoyagi, I. Iswandi, M. Kim, J.-I. Takada, K. Hamaguchi, and R. Kohno. Body motion and channel response of dynamic body area channel. In *Proceedings of the 5th European Conference on Antennas and Propagation*, (EUCAP), pages 3138–3142, 2011.

[4] I. Armenti, P. Asare, J. Su, and J. Lach. A methodology for developing quality of information metrics for body sensor design. In *Wireless Health*, 2012.

[5] D. Arney, M. Pajic, J. M. Goldman, I. Lee, R. Mangharam, and O. Sokolsky. Toward patient safety in closed-loop medical device systems. In *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems*, ICCPS '10, pages 139–148, 2010.

[6] P. Asare, R. F. Dickerson, X. Wu, J. Lach, and J. A. Stankovic. Bodysim: a multi-domain modeling and simulation framework for body sensor networks research and design. In *Proceedings of the 8th International Conference on Body Area Networks*, pages 177–180. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013.

[7] P. Asare, R. F. Dickerson, X. Wu, J. Lach, and J. A. Stankovic. Bodysim: a multi-domain modeling and simulation framework for body sensor networks research and design. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor*

*Systems*, pages 177–180. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013 (Demo).

[8] P. Asare, J. Lach, and J. A. Stankovic. FSTPA-I: A formal approach to hazard identification via system theoretic process analysis. In *Proceedings of the 4th ACM/IEEE International Conference of on Cyber-Physical Systems*, April 2013.

[9] P. Asare, J. Lach, J. A. Stankovic, Y. Zhang, P. L. Jones, and S. Weininger. Towards a framework for safety analysis of body sensor networks. In *ICST International Conference on Body Area Networks*, 2013.

[10] A. D. Association. Standards of medical care in diabetes 2014. *Diabetes Care*, 37(Supplement 1):S14–S80, 2014.

[11] A. Banerjee, S. Kandula, T. Mukherjee, and S. K. S. Gupta. Band-aide: A tool for cyber-physical oriented analysis and design of body area networks and devices. *ACM Trans. Embed. Comput. Syst.*, 11(S2):49:1–49:29, Aug 2012.

[12] A. Banerjee, S. Verma, P. Bagade, and S. K. Gupta. Health-dev: Model based development pervasive health monitoring systems. In *Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on*, pages 85–90. IEEE, 2012.

[13] J. H. Barnett. Applications of Boolean Algebra: Claude Shannon and circuit design. *Loci*, July 2013.

[14] N. Baysal, F. Cameron, B. A. Buckingham, D. M. Wilson, H. P. Chase, D. M. Maahs, B. W. Bequette, B. A. Buckingham, D. M. Wilson, T. Aye, and et al. A novel method to detect pressure-induced sensor attenuations (pisa) in an artificial pancreas. *Journal of Diabetes Science and Technology*, 8(6):1091âĂŞ1096, Oct 2014.

[15] I. Beretta, F. Rincon, N. Khaled, P. R. Grassi, V. Rana, and D. Atienza. Design exploration of energy-performance trade-offs for wireless sensor networks. In *Proceedings of the 49th Annual Design Automation Conference*, DAC '12, pages 1043–1048, New York, NY, USA, 2012. ACM.

[16] Blender Foundation. Blender. http://www.blender.org/.

[17] J. C. Boyd and D. E. Bruns. Quality specifications for glucose meters: Assessment by simulation modeling of errors in insulin dose. *Clinical Chemistry*, 47(2):209–214, 2001.

[18] M. Breton and B. Kovatchev. Analysis, modeling, and simulation of the accuracy of continuous glucose sensors. *Journal of diabetes science and technology (Online)*, 2(5):853–862, Sept. 2008.

[19] M. D. Breton. Physical activity: The major unaccounted impediment to closed loop control. *Journal of diabetes science and technology (Online)*, 2(1):169–174, Jan. 2008.

[20] M. D. Breton, S. A. Brown, C. H. Karvetski, L. Kollar, K. A. Topchyan, S. M. Anderson, and B. P. Kovatchev. Adding heart rate signal to a control-to-range artificial pancreas system improves the protection against hypoglycemia during exercise in type 1 diabetes. *Diabetes Technology & Therapeutics*, 16(8):506–511, Apr. 2014.

[21] B. Calhoun, J. Lach, J. Stankovic, D. Wentzloff, K. Whitehouse, A. Barth, J. Brown, Q. Li, S. Oh, N. Roberts, and Y. Zhang. Body sensor networks: A holistic approach from silicon to users. *Proceedings of the IEEE*, 100(1):91 –106, jan. 2012.

[22] C. Campos. Chronic hyperglycemia and glucose toxicity: pathology and clinical sequelae. *Postgraduate Medicine*, November 2012.

[23] R. H. S. Carpenter. Homeostasis: a plea for a unified approach. *AJP: Advances in Physiology Education*, 28(4):180–187, Dec. 2004.

[24] Centers for Disease Control and Prevention. Body mass index. [Online]: http://www.cdc.gov/healthyweight/assessing/bmi/index.html.

[25] S. Chen, A. T. Barth, J. T. Barth, B. C. Bennett, M. Brandt-Pearce, D. K. Broshek, J. R. Freeman, H. L. Samples, and J. Lach. Aiding diagnosis of normal pressure hydrocephalus with enhanced gait feature separability. In *Proceedings of the conference on Wireless Health*, WH '12, pages 3:1–3:8, New York, NY, USA, 2012. ACM.

[26] S. Chen, C. L. Cunningham, B. C. Bennett, and J. Lach. Enabling longitudinal assessment of ankle-foot orthosis efficacy for children with cerebral palsy. In *Pro-*

*ceedings of the 2nd Conference on Wireless Health*, WH '11, pages 4:1–4:10, New York, NY, USA, 2011. ACM.

[27] S. Chen, J. Lach, O. Amft, M. Altini, and J. Penders. Unsupervised activity clustering to estimate energy expenditure with a single body sensor. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6, 2013.

[28] E. M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT, 1999.

[29] W. L. Clarke, S. Anderson, M. Breton, L. K. S. Patek, and B. Kovatchev. Closed-loop artificial pancreas using subcutaneous glucose sensing and insulin delivery, and a model-predictive control algorithm: The virginia experience. *Journal of Diabetes Science and Technology*, 3(5):1031–1038, 2009.

[30] P. E. Cryer. *Hypoglycemia: Pathophysiology, diagnosis, and treatment*. Oxford University Press, New York, 1997.

[31] V. De Santis, M. Feliziani, and F. Maradei. Safety assessment of uwb radio systems for body area network by the method. *Magnetics, IEEE Transactions on*, 46(8):3245–3248, 2010.

[32] S. Delp, F. Anderson, A. Arnold, P. Loan, A. Habib, C. John, E. Guendelman, and D. Thelen. Opensim: Open-source software to create and analyze dynamic simulations of movement. *Biomedical Engineering, IEEE Transactions on*, 54(11):1940–1950, 2007.

[33] A. Derhab, F. Ounini, and B. Remli. Mob-tossim: An extension framework for tossim simulator to support mobility in wireless sensor and actuator networks. *Distributed Computing in Sensor Systems and Workshops, International Conference on*, 0:300–305, 2012.

[34] e2v Technologies. Mics-2610 $O_3$ sensor datasheet, 2008. [Online]: http://www.cdiweb.com/datasheets/e2v/mics-2610.pdf [Last Accessed]: April 10, 2015.

[35] J. Eker, J. Janneck, E. Lee, J. Liu, X. Liu, J. Ludvig, S. Neuendorffer, S. Sachs, and Y. Xiong. Taming heterogeneity - the ptolemy approach. *Proceedings of the IEEE*, 91(1):127 – 144, Jan. 2003.

[36] FARO®. FARO laser scanner Focus$^{3D}$ X series. [Online]: http://www.faro.com/products/3d-surveying/laser-scanner-faro-focus-3d/overview.

[37] C. H. Fleming. *Safety-driven early concept analysis and develpment*. PhD thesis, Massachussetts Institute of Technology, 2015.

[38] GE Medical Systems Inforamtion Technologies. *Dash 3000/4000/5000 Patient Monitor Operator's Manual. 2023896-026 Revision A*. 2005.

[39] J. Gleick. *The information: a history, a theory, a flood*. Pantheon Books, 2011.

[40] J. Gong, P. Asare, J. Lach, and Y. Qi. Piecewise linear dynamical model for actions clustering from inertial body sensors with considerations of human factors.

[41] R. A. Harrigan, T. C. Chan, and W. J. Brady. Electrocardiographic electrode misplacement, misconnection, and artifact. *The Journal of Emergency Medicine*, 43(6):1038 – 1044, 2012.

[42] H. W. Heinrich. *Industrial Accident Prevention: A scientific approach*. McGraw-Hill, 1931.

[43] T. Hermitte. D5.9 Review of accident causation models used in road accident research of the EC FP7 project DaCoTA. Technical report, European Commission Directorate General for Mobility and Transport, 2012.

[44] R. Hester, A. Brown, L. Husband, R. Iliescu, W. A. Pruett, R. L. Summers, and T. Coleman. Hummod: A modeling environment for the simulation of integrative human physiology. *Frontiers in Physiology*, 2(12), 2011.

[45] E. Hollnagel. *Barriers and Accident Prevention*. Ashgate, 2004.

[46] Z. Jiang, M. Pajic, and R. Mangharam. Model-based closed-loop testing of implantable pacemakers. In *Cyber-Physical Systems (ICCPS), 2011 IEEE/ACM International Conference on*, pages 131–140, 2011.

[47] B. Kim, A. Ayoub, O. Sokolsky, I. Lee, P. Jones, Y. Zhang, and R. Jetley. Safety-assured development of the GPCA infusion pump software. In *Proceedings of ACM International Conference on Embedded Software*, EMSOFT '11, pages 155 –164, October 2011.

[48] A. E. Kitabchi, G. E. Umpierrez, and R. A. Kreisberg. Hyperglycemic crises in adult patients with diabetes: a consensus statement from the American Diabetes Association. *Diabetes Care*, 29(12), Dec 2006.

[49] B. P. Kovatchev, M. Breton, C. DallaMan, and C. Cobelli. *In silico* preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *Journal of diabetes science and technology*, 3(1):44–55, 2009.

[50] B. P. Kovatchev, D. J. Cox, L. A. Gonder-Frederick, D. Young-Hyman, D. Schlundt, and W. Clarke. Assessment of risk for severe hypoglycemia among adults with iddm: validation of the low blood glucose index. *Diabetes Care*, 21(11):1870–1875, Nov 1998.

[51] B. P. Kovatchev, E. Otto, D. Cox, L. Gonder-Frederick, and W. Clarke. Evaluation of a new measure of blood glucose variability in diabetes. *Diabetes Care*, 29(11):2433–2438, Nov 2006.

[52] B. P. Kovatchev, S. D. Patek, E. A. Ortiz, and M. D. Breton. Assessing sensor accuracy for non-adjunct use of continuous glucose monitoring. *Diabetes Technology & Therapeutics*, 17(3):177–186, Dec. 2014.

[53] K. Kroenke, R. Spitzer, and J. Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[54] K. Kroenke and R. L. Spitzer. The phq-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9):509–515, 2002.

[55] E. A. Lee and P. Varaiya. *Structure and Interpretation of Signals and Systems*. LeeVaraiya.org, 2 edition, 2011.

[56] H. Lee, A. Cerpa, and P. Levis. Improving wireless simulation through noise modeling. In *Proceedings of the 6th international conference on Information processing in sensor networks*, IPSN '07, pages 21–30, New York, NY, USA, 2007. ACM.

[57] I. Lee, G. J. Pappas, R. Cleaveland, J. Hatcliff, B. H. Krogh, P. Lee, H. Rubin, and L. Sha. High-confidence medical device software and systems. *Computer*, 39(4):33–38, April 2006.

[58] N. Leveson. A new approach to hazard analysis for complex systems. In *Int. Conference of the System Safety Society*, August 2003.

[59] N. Leveson. A new accident model for engineering safer systems. *Safety Science*, 42(4):237 – 270, 2004.

[60] N. Leveson. *Engineering a safer world: systems thinking applied to safety*. Engineering systems. MIT Press, Cambridge, Mass, 2011.

[61] N. Leveson, M. Couturier, J. Thomas, M. Dierks, D. Wierz, B. Psaty, and S. Finkelstein. Applying system engineering to pharmaceutical safety. *Journal of Healthcare Engineering (to appear)*. [Online] http://sunnyday.mit.edu/papers/healthcare-eng-final.doc.

[62] P. Levis, N. Lee, M. Welsh, and D. Culler. Tossim: accurate and scalable simulation of entire tinyos applications. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, SenSys '03, pages 126–137, New York, NY, USA, 2003. ACM.

[63] N. Li, J. Cruz, C. S. Chien, S. Sojoudi, B. Recht, D. Stone, M. Csete, D. Bahmiller, and J. C. Doyle. Robust efficiency and actuator saturation explain healthy heart rate control and variability. *Proceedings of the National Academy of Sciences*, 111(33):E3476–E3485, 2014.

[64] L. Magni, D. M. Raimondo, C. D. Man, M. Breton, S. Patek, G. D. Nicolao, C. Cobelli, and B. P. Kovatchev. Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis. *Journal of diabetes science and technology (Online)*, 2(4):630–635, July 2008.

[65] C. Man, R. Rizza, and C. Cobelli. Meal simulation model of the glucose-insulin system. *Biomedical Engineering, IEEE Transactions on*, 54(10):1740–1749, Oct 2007.

[66] J. C. Maxwell. On governors. *Proceedings of the Royal Society of London*, 16:270–283, 1867.

[67] Mayo Clinic. Radiation therapy: Risks. [Online]: http://www.mayoclinic.org/tests-procedures/radiation-therapy/basics/risks/prc-20014327 [Last Updated: July 26, 2014].

[68] K. C. McCowen, A. Malhotra, and B. R. Bistrian. Stress-induced hyperglycemia. *Critical Care Clinics*, 17(1):107 – 124, 2001.

[69] A. Murugesan, O. Sokolsky, S. Rayadurgam, M. Whalen, M. Heimdahl, and I. Lee. Linking abstract analysis to concrete design: A hierarchical approach to verify medical cps safety. In *Cyber-Physical Systems (ICCPS), 2014 ACM/IEEE International Conference on*, pages 139–150, April 2014.

[70] National Cancer Institute. Tumor grade. [Online]: http://www.cancer.gov/cancertopics/diagnosis-staging/prognosis/tumor-grade-fact-sheet. [Last Updated]: May 3, 2013.

[71] National Heart, Lung, and Blood Institute. What are the health risks of overweight and obesity? [Online]: http://www.nhlbi.nih.gov/health/health-topics/topics/obe/risks.html.

[72] National ICT Australia (NICTA). Castalia wireless sensor network simulator. http://castalia.research.nicta.com.au/index.php/en/.

[73] National Institute of Diabetes and Digestive and Kidney Diseases. Your guide to diabetes: Type 1 and type 2. [Online]: http://diabetes.niddk.nih.gov/dm/pubs/type1and2/index.aspx [Last Updated: February 12, 2014].

[74] National Institute of Standards and Technology. Fire dynamics, 2010, 2013. [Online]: http://www.nist.gov/fire/fire_behavior.cfm [Last Updated]: July 16, 2013. [Last Accessed]: April 10, 2014.

[75] B. W. Nugent. Hyperosmolar hyperglycemic state. *Emergency Medicine Clinics of North America*, 23(3):629 – 648, 2005. Endocrine and Metabolic Emergencies Endocrine and Metabolic Emergencies.

[76] K. O'Donovan, B. Greene, D. McGrath, R. O'Neill, A. Burns, and B. Caulfield. Shimmer: A new tool for temporal gait analysis. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 3826–3829, 2009.

[77] B. D. Owens, M. S. Herring, N. Dulac, N. Leveson, M. Ingham, , and K. A. Weiss. Application of a safety-driven design methodology to an outer planet exploration mission. In *IEEE Aerospace Conference*, 2008.

[78] M. Pajic, R. Mangharam, O. Sokolsky, D. Arney, J. Goldman, and I. Lee. Model-driven safety analysis of closed-loop medical systems. *IEEE Transactions on Industrial Informatics*, 10(1):3–16, Feb 2014.

[79] A. Pantelopoulos and N. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(1):1 –12, Jan. 2010.

[80] A. Parnandi, E. Wade, and M. Mataric. Motor function assessment using wearable inertial sensors. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 86–89, 2010.

[81] S. D. Patek, B. W. Bequette, M. Breton, B. A. Buckingham, E. Dassau, F. J. Doyle, J. Lum, L. Magni, and H. Zisser. In silico preclinical trials: Methodology and engineering guide to closed-loop control in type 1 diabetes mellitus. *Journal of Diabetes Science and Technology*, 3(2):269–282, Mar. 2009.

[82] S. D. Patek, E. A. Ortiz, L. Farhy, J. M. Lobo, J. Isbell, J. L. Kirby, and A. McCall. Population-specific models of glycemic control in intensive care: Towards a simulation-based methodology for protocol optimization. In *IEEE American Control Conference*, 2015.

[83] S. Patel, R. Hughes, T. Hester, J. Stein, M. Akay, J. Dy, and P. Bonato. Tracking motor recovery in stroke survivors undergoing rehabilitation using wearable technology. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 6858–6861, 2010.

[84] Providence Heart and Vascular Institute Portland Protocol (Version 2008.1). Portland continuous intravenous insulin protocol. [Online]: http://oregon.providence.org/our-services/p/portland-diabetes-project/research/the-portland-protocol [Last accessed: March 18, 2015].

[85] J. Pukite and P. Pukite. *Modeling for Reliability Analysis: Markov Modeling for Reliability, Maintainability, Safety, and Supportability Analyses of Complex Systems*. Wiley-IEEE Press, 1998.

[86] Python Software Foundation. Python programming language. http://python.org/.

[87] J. Rasmussen. Risk management in a dynamic society: a modelling problem. *Safety Science*, 27(2âĂŞ3):183 – 213, 1997.

[88] J. T. Reason. *Managing the Risks of Organisational Accidents*. Ashgate, 1997.

[89] W. T. Riley, D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, and R. Mermelstein. Health behavior models in the age of mobile interventions: are our theories up to the task? *Translational behavioral medicine*, 1(1):53?71, March 2011.

[90] Riverbed. Opnet. [Online]: http://www.riverbed.com/products/opnet.html.

[91] G. Schöner. Dynamical systems approaches to cognition. *Cambridge handbook of computational cognitive modeling*, pages 101–126, 2008.

[92] J. M. Schumann. Automated theorem proving in software engineering. 2001.

[93] C. E. Shannon. A symbolic analysis of relay and switching circuits. Master's thesis, Massachusetts Institute of Technology, 1940.

[94] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[95] H. A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 3 edition, 1996.

[96] R. L. Spitzer, J. B. Williams, and K. Kroenke. The patient health questionnaire-9. [Online]: http://www.phqscreeners.com/pdfs/02_PHQ-9/English.pdf [Last Accessed]: April 6, 2015.

[97] D. H. Stamatis. *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. American Society for Quality, 1995.

[98] G. M. Steil, D. Deiss, J. Shih, B. Buckingham, S. Weinzimer, and M. S. Agus. Intensive care unit insulin delivery algorithms: Why so many? how to choose? *Journal of Diabetes Science and Technology*, 3(1):125–140, 2009.

[99] D. J. Stone, L. A. Celi, and M. Csete. Engineering control into medicine. *Journal of Critical Care*, (0):–, 2015.

[100] Synertial. Anamizoo IGS 180i motion capture system. [Online]: http://www.synertial.com/products/180i.

[101] S. Tepsuporn. BodySim: A simulation framework for body sensor networks, 2015.

[102] The MathWorks. Matlab/simulink. http://www.mathworks.com/products/simulink/.

[103] The MathWorks®. Matlab/simulink. http://www.mathworks.com/products/matlab/.

[104] The New England Chapter of the System Safety Society. System safety: A science and technology primer, April 2002.

[105] J. Thomas. *Extending and Automating a Systems-Theoretic Hazard Analysis for Requirements Generation and Analysis*. PhD thesis, Massachussetts Institute of Technology, 2013.

[106] Y. Toft, G. Dell, K. K. Klockner, and A. Hutton. Models of causation: Safety. In *The Core Body of Knowledge for Generalist OHS Professionals*, HaSPA (Health and Safety Professionals Alliance). Safety Institute of Australia, Tullamarine, VIC, 2012.

[107] S. Tschirner, L. Xuedong, and W. Yi. Model-based validation of qos properties of biomedical sensor networks. In *Proceedings of the 8th ACM international conference on Embedded software*, EMSOFT '08, pages 69–78, New York, NY, USA, 2008. ACM.

[108] U.S. Environmental Protection Agency. Health effects of ozone in patients with asthma and other chronic respiratory diseases, 2015. [Online]; http://www.epa.gov/apti/ozonehealth/effects.htm [Last Updated]: January 30, 2015.

[109] U.S. Food and Drug Administration. Virtual family. [Online]: http://www.fda.gov/AboutFDA/CentersOffices/

OfficeofMedicalProductsandTobacco/CDRH/CDRHOffices/ucm302074.htm [Last Accessed]: Aptil 10, 2014.

[110] U.S. Food and Drug Administration. What we do. http://www.fda.gov/aboutfda/whatwedo/.

[111] G. Van den Berghe, P. Wouters, F. Weekers, C. Verwaest, F. Bruyninckx, M. Schetz, D. Vlasselaers, P. Ferdinande, P. Lauwers, and R. Bouillon. Intensive insulin therapy in critically ill patients. *New England Journal of Medicine*, 345(19):1359–1367, 2001. PMID: 11794168.

[112] J. Ventura and K. Chowdhury. Markov modeling of energy harvesting body sensor networks. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, pages 2168–2172, 2011.

[113] W. Vesely, F. Goldberg, N. Roberts, and D. Haasl. Fault tree handbook (NUREG-0492). Technical report, U.S. Nuclear Regulatory Commission, 1981.

[114] J. G. Webster and J. W. Clark, editors. *Medical instrumentation: application and design*. John Wiley & Sons, Hoboken, NJ, 4th ed edition, 2010.

[115] B. J. West and N. Scafetta. Nonlinear dynamical model of human gait. *Phys. Rev. E*, 67:051917, May 2003.

[116] N. Wiener. *Cybernetics: Or the Control and Communication in the Animal and the Machine*. The MIT Press, 2 edition, 1965.

[117] D. F. Williams. On the mechanisms of biocompatibility. *Biomaterials*, 29(20):2941 – 2953, 2008.

[118] World Health Organization and International Diabetes Federation. *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: a report of a WHO/IDF consultation*. WHO Document Production Services, 2006.

[119] A. Young, M. Ling, and D. K. Arvind. IMUSim: A simulation environment for inertial sensing algorithm design and evaluation. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 199–210, 2011.

[120] A. D. Young, M. J. Ling, and D. K. Arvind. Orient-2. *Proceedings of the 4th workshop on Embedded Networked Sensors - EmNets '07*, 2007.

[121] Y. Zigel, A. Cohen, and A. Katz. The weighted diagnostic distortion (WDD) measure for ECG signal compression. *Biomedical Engineering, IEEE Transactions on*, 47(11):1422–1430, Nov. 2000.