Towards Predictive Medicine — On Remote Monitoring, Privacy and Scientific Bias

Daniel Aranki

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-145
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-145.html

August 11, 2017

Acknowledgement

See dissertation.

Towards Predictive Medicine – On Remote Monitoring, Privacy and Scientific Bias

by
Daniel Aranki

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Computer Science
and the Designated Emphasis
in
Communication, Computation, and Statistics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Ruzena Bajcsy, Chair
Professor John F. Canny
Professor Deirdre Mulligan

Summer 2017

Towards Predictive Medicine – On Remote Monitoring, Privacy and Scientific Bias

Abstract

Towards Predictive Medicine – On Remote Monitoring, Privacy and Scientific Bias

by

Daniel Aranki

Doctor of Philosophy in Computer Science

and the Designated Emphasis

in Communication, Computation, and Statistics

University of California, Berkeley

Professor Ruzena Bajcsy, Chair

The current healthcare model in the United States of America (US) is reactive in nature. That is, individuals usually seek medical attention after symptoms manifest. In 2015, the total cost of healthcare in the US was $3.2 trillion (17.8% of US Gross Domestic Product), which amounts to $9,990 per capita. In the same year, the 30-days all-condition rate of unplanned rehospitalizations in patients in the Medicare fee-for-service program was around 17.9%; and between October 1, 2003 and December 31, 2003, 3.5% of patients in the same program died within 30 days of initial discharge.

Alternatively, a healthcare model that utilizes medical intervention based on personalized predictions of the patient's clinical status and possible deterioration could potentially decrease costs, unplanned rehospitalizations and mortality rates. This model also has the potential to improve the overall quality of care. We refer to this model as the predictive healthcare model.

In this dissertation, we examine three outstanding challenges towards fully realizing the predictive healthcare model as the prevalent care model. Namely, i) we investigate means to streamline the costly longitudinal epidemiological studies using remote mobile monitoring and introduce the Berkeley Telemonitoring project; ii) we investigate the privacy challenge that is particular to the remote monitoring model and introduce the Private Disclosure of Information (PDI) semantic privacy model; and iii) we investigate the problem of publication bias in empirical sciences (including biomedicine) that hinders the credibility of empirical scientific findings and introduce a statistical test that detects bias in a sample of scientific publications which utilize the Student t-test.

To All Those Who Are Denied Pursuing Their Dreams

Acknowledgements

I have collected remarkable amounts of debt to many people during and before my doctoral studies. On academic, professional and personal levels, I owe a great deal of gratitude to my advisor, Prof. Ruzena Bajcsy, who took me under her wing and advised me every step along my journey at UC Berkeley. She has been an advisor and a mentor, and I am truly honored to have worked with her all these years. Her continued support made this dissertation possible: from hashing out the ideas discussed hereinafter, to finding the resources for my doctoral work, and beyond. Her incredible experience and knowledge is surpassed only by the quality of the time spent in her company and her hospitality; and the amount of enrichment she has blessed me with.

In many ways I consider Prof. John F. Canny my second academic advisor. His vast knowledge and experience are unparalleled, and I feel privileged to have worked with him. He taught me plenty, and provided feedback and pieces of advice on many aspects of my work, including the detection and estimation of publication bias in empirical research.

Working in the technological realm, I sometimes fall in the trap of confining myself to a narrow point of view. Prof. Deirdre Mulligan has always been there to widen my horizons and unblock my academic vision from the curse of confinement. Her broad understanding of privacy is truly remarkable; so much of what I know about privacy is due to her teaching.

The feedback by Professors Bajcsy, Canny and Mulligan, both during my doctoral work and during the writing of this dissertation, led to countless improvements in the quality of this manuscript and the arguments made within. Any mistakes or typographical errors that made it to the finished dissertation are my sole responsibility, not theirs.

One cannot embark on the journey of doctoral studies without support from many people beyond the academic advisor. In my case, I was blessed with a group of exceptional people who supported and collaborated with me, empowering me through this endeavor. I would like to thank Gregorij Kurillo for his continued support, particularly in those uncountable times when he made himself available in weekends and vacations; he certainly was and continues to be a pillar of support. I owe a great deal to Ferda Ofli, a true friend of rare kind, who not only shared his office space with me during my first years at Berkeley, but advised me along the way. I learned a great deal about medicine and medical research design from David M. Liebovitz, MD and the Northwestern Medical Faculty Foundation team, I am in their debt. To all the wonderful people who ever set foot in the Human-Assistive Robotic Technologies (HART) lab (formerly known as Teleimmersion Lab), particularly Victor Shia, Aaron Bestick, Katie Driggs-Campbell, and Robert Matthew: thank you from the bottom of my heart, you made my journey much more pleasant than it would've been otherwise. Many thanks to Roel Dobbe, Jaime Fernández Fisac, and Cathy Wu, whose stimulating discussions about the future of automation and Artificial Intelligence (AI) research and their societal implications enriched my worldview. Much of my vision in writing this dissertation became clearer due to the rest of my qualifying exam committee, Professors Doug Tygar and Thomas Courtade; I am grateful for their support and advice. I owe more gratitude than I could repay to the late Aimee Tabor, who passed away way too young; she will always be remembered. Navigating the administrative requirements of a doctoral degree can be overwhelming at times, and in more instances than I can remember, the indispensable Electrical Engineering and Computer Sciences (EECS) department staff assisted me in ways that cannot be replaced; I want to particularly acknowledge Jessica Gamble, Shirley Salanio and Audrey Sillers.

Much of the work described here would not have been possible without the effort The Berkeley Telemonitoring Project team.[1] Explicitly, I would like to thank Posu Yan, Arjun Chopra, Eugene Song, Adarsh Mani, Phillip Azar, Jochem van Gaalen, Quan Peng, Priyanka Nigam, Maya P. Reddy, Sneha Sankavaram, Qiyin Wu, Uma Balakrishnan, Hannah Sarver, Lucas Serven, Carlos Asuncion, Kaidi (Kate) Du, Caitlin Gruis, Gao Xian Peh, Yu (Sean) Xiao and Joany Gao.

I acquired invaluable knowledge from Professor Jim Pitman, not only in the classroom, but also in our chats about the formal methods in detecting publication bias. His input helped shape the statistical test for detecting publication bias to what it is today. For all of that, I am greatly thankful.

The research presented in this dissertation was made possible by the generous awards and funding by the National Science Foundation (NSF), the Department of Health and Human Services (HHS), and others. In particular, I would like to acknowledge The Team for Research in Ubiquitous Secure Technology (TRUST) (NSF award number CCF-0424422), Strategic Health IT Advanced Research Projects (SHARP) (Grant Number HHS 90TR0003/01), and Center for Long-Term Cybersecurity (CLTC).

An integral part of my doctoral work has been teaching, an activity that I enjoy immensely. I would like to thank Professors Anant Sahai and Ali Niknejad for providing me the amazing opportunity to build and teach EE16A with them in its first scaled offering, Fall 2015. The team of EE16A in Fall 2015 is responsible for much of my growing as a teacher and a curriculum developer. I have truly never taught with a group of more professional, dedicated and talented people. In addition to the gratitude that I owe the whole team, I would like to explicitly thank Filip Maksimovic, Vidya Muthukumar, Claire Lochner, Preetum Nakkiran, Daniel Calderone, Paul Rigge and Kene Akimaluto. I've had the privilege to teach EE16A again, in lecturing capacity, in Summer 2017. For Summer 2017, I'd like to thank the entire staff of EE16A, in particular my co-lecturers Filip Maksimovic and Vasuki Narasimha Swamy for supporting me while I was making the finishing touches in this dissertation.

Life is full of people who fill our hearts with warmth and joy, friends who are there every step of the way. I am privileged to have traveled this journey accompanied by many of these. I would like to thank my friends for the countless memories and the tireless support. Shout-outs to Yusuf Bugra Erol, Hélène Viart, Fabien Chraim, Daniel Calderone (yes, the same Dan from EE16A), Sebastian Benthall, Rondu Gantt, Gus and Georgina Totah, Micheal Khayat, Rami Issa, Amen Shihada, Mohammad Nassar, Hani Ayoub, Ibrahem Baransi and Doreen Danial.

Many thanks to my one and only, Karen Seif, who never ceased to support me. You are a beacon of light that shines even in the brightest of days. In all truthfulness, this dissertation would not have been near possible if it hasn't been to your support and care.

Last but not least, I wish to extend my bottomless gratitude to my family, both immediate and extended. First and foremost, to my parents, Rose and Issam Aranki, who not only had sacrificed repeatedly to raise me, but have always believed in me–even when I doubted myself. To my sister, Grace Shaheen, who spent many early mornings and late nights patiently listening to my frustrated stories about lack of research progress; and to her husband, Louis Shaheen, for without their support during my undergraduate days, I would not have even made it to Berkeley. To my brother, Father Eugenios (Ευγένιοζ) Aranki, who has been a source of unlimited personal support and wisdom. To my sister, Carole Naser, whose kind heart is of rarefied uniqueness. To my cousin, Mansoor Zaknoon, who has always painted a smile on my face, specially when I needed it most. To my cousin, Shereen Naser, for all the stimulating conversations she has engaged me

[1] `https://telemonitoring.berkeley.edu/team/`

with. Finally, to my aunt, Ghada Saba, who has been my family away from my family.

The views and opinions expressed in this dissertation are mine, and do not necessarily reflect the views and opinions of any of the people, entities or agencies listed above.

Contents

1 Introduction **1**
1.1 Genesis . . . 2
1.1.1 Readmission Rates . . . 4
1.1.2 Mortality Rates . . . 6
1.1.3 Cost of Healthcare . . . 7
1.2 The Gaps Towards Predictive Medicine . . . 8
1.2.1 The Need for Reliable Data . . . 10
1.2.2 The Need for Reliable Science . . . 11
1.2.3 Privacy in Predictive Healthcare . . . 12
1.3 Main Contributions . . . 13
1.4 Dissertation Organization . . . 13

2 Telemonitoring for Predctive Medicine **18**
2.1 Introduction . . . 19
2.2 Feasibility and Related Work . . . 20
2.2.1 Telemonitoring in Patients with CHF . . . 20
2.2.2 Mobile Health (mHealth) . . . 22
2.2.3 Quantitative Measurement of Physical Activity . . . 23
2.3 Intervention . . . 24
2.4 CHF Study . . . 24
2.4.1 Introduction . . . 24
2.4.2 Collected Data . . . 26
2.4.3 Study Findings . . . 27
2.4.4 Challenges and Lessons Learned . . . 28
2.5 Commercial Solution . . . 30
2.5.1 Introduction . . . 30
2.5.2 Samsung Digital Health and S Health . . . 30
2.5.3 Apple Health, ResearchKit, and CareKit . . . 31
2.5.4 Google Fit . . . 32
2.5.5 Other Commercial Platforms . . . 33
2.6 Challenges . . . 33
2.7 The Berkeley Telemonitoring Project . . . 36
2.7.1 Introduction . . . 36
2.7.2 Framework Structure . . . 38
2.8 RunningCoach Study . . . 49
2.8.1 The Premise . . . 49
2.8.2 System and Study Design . . . 49

2.9 Summary and Discussion . . . 51

3 Privacy in Telemonitoring **58**
3.1 Introduction . . . 59
3.2 User Privacy and Acceptability of Telemonitoring . . . 60
3.2.1 Introduction . . . 60
3.2.2 CHF – Privacy and Acceptability Study . . . 61
3.2.3 RunningCoach – Privacy and Acceptability Study . . . 63
3.2.4 What's Next? . . . 66
3.3 The Inference Design Principle . . . 66
3.3.1 Existing Privacy Design Principles and Practices . . . 66
3.3.2 Our Philosophical View on Privacy – The Inference Threat . . . 76
3.4 Private Disclosure of Information (PDI) . . . 77
3.4.1 Introduction . . . 77
3.4.2 Problem Formulation . . . 78
3.4.3 Further Analysis and PDI Properties . . . 83
3.4.4 Learning the Privacy Mapping Function . . . 87
3.4.5 Experimentation . . . 89
3.5 Summary and Discussion . . . 93

4 A Data-Driven Approach to Detecting Publication Bias **98**
4.1 Preface – Thought Experiment . . . 99
4.1.1 Selection Bias . . . 100
4.2 Introduction . . . 101
4.2.1 Predictive Medicine . . . 102
4.2.2 Significance Analysis – Setting . . . 102
4.2.3 Publication Bias in General . . . 104
4.2.4 Chapter Organization and Contributions . . . 105
4.3 Related Work . . . 105
4.4 Formal Method . . . 106
4.5 Implementation . . . 109
4.5.1 Computing $\mathbb{E}\exp(\theta X(n,e))$ . . . 110
4.5.2 Partitioning the Degrees of Freedom (DOFs) . . . 111
4.6 Experiment . . . 114
4.7 Summary . . . 117

5 Final Thoughts **121**

Glossary **124**

Bibliography **128**

Appendices **140**
.1 Appendix I – Publication Bias Data . . . 141

List of Figures

1.1 Age-adjusted mortality rates for prostate cancer (men) and breast cancer (women) between the years 1968 and 2015. The data were obtained from the Center for Disease Control and Prevention (CDC) [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015]. . . . 3
1.2 (a) The general workflow of the current healthcare model; and (b) the general workflow in the predictive healthcare model. . . . 4
1.3 Medicare 30-Day, all-condition hospital readmission rate. Data were provided by the Centers for Medicare & Medicaid Services (CMS) in written communication to President Obama. The plotted series reflects a 12-month moving average of the hospital readmission rates reported for discharges occurring in each month [Obama, 2016]. . . . 5
1.4 Age-adjusted mortality rates for i) deaths not caused by external causes; and ii) deaths caused by congestive heart failure (CHF) in men and women between the years 1968 and 2015 in the United States of America (US). The data were obtained from the CDC [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015]. . . . 7
1.5 The national health expenditure in the US (total and per capita) since 1960, adjusted to 2015 United States Dollars (USD). . . . 8
1.6 The proportion of the US national health expenditure from the gross domestic product (GDP), since 1960. . . . 9

2.1 The general workflow in the predictive healthcare model. This model also coincides with the workflow of telemonitoring systems. . . . 20
2.2 (a) The architecture of the CHF study system; (b) the CHF study dashboard main menu; (c) The CHF telemonitoring app; (d) minute-by-minute energy expenditures (EE) estimates, overlaid by the classified state of the phone; and (e) daily average EE estimates, bars depict daily averages while the line depicts moving 7-day averages. 25
2.3 Excerpts of battery recharging and consumption patterns of (a) subject 1012 over the course of three days; and (b) subject 1013 over the course of 17 days. . . . 29
2.4 (a) A screenshot of the Samsung Health summary screen; (b) a screenshot of the Apple Health activity tracking screen; and (c) a screenshot of the Google Fit activity tracking screen. . . . 31
2.5 The architectural breakdown of the Berkeley Telemonitoring framework. . . . 40
2.6 (a) A screenshot of Bluetooth/Bluetooth Low Energy (BLE) stack scanning for nearby devices; (b) a demonstration of the finger-based heart rate estimator (top: an external photo of the setting, bottom: the interface in the telemonitoring app); and (c) a demonstration of the face-based heart rate estimator. . . . 43

2.7 (a) The different components of the survey; (b) an example of a text question/answer survey node; and (c) an example of a text question and radio answer survey node. . 45
2.8 The finite state machine description of the TI protocol. 46
2.9 Screenshots from the RunningCoach app: (a) the runner's physical parameters screen; (b) an example cadence training regimen; and (c) the app's home screen. . 50
2.10 Screenshots from the RunningCoach app: (a) the screen that is displayed during the run; (b) a sample post-run survey question; and (c) a sample post-run summary. 51
2.11 Sample plots from the server dashboard: (a) cadence plot; (b) distance-covered plot; (c) EE plot; (d) heart-beat rate plot; (e) speed plot; and (f) GPS plot showing the path of run 22. 52

3.1 Pre-study acceptability survey results for the CHF study (N=15). 61
3.2 The acceptability survey results, in pre- and post-study surveys, for the 5 subjects who completed the CHF post-study survey. 62
3.3 The subjects' responses to the pre-CHF-study survey privacy-related questions administered, inquiring about how comfortable they are when sharing their data on weight, physical activity levels, location, and blood pressure ($N = 15$). 63
3.4 Post-study acceptability survey results for the RunningCoach study (N=6). 64
3.5 The responses to privacy related question administered after to the RunningCoach study, inquiring about how comfortable the users are sharing their data on weight, physical activity levels, location, and heart-beat rate values (N=6). 65
3.6 The Private Disclosure of Information (PDI) threat model: a statistical inference attack by a passive eavesdropper. 79
3.7 The statistical graphical model of PDI. S and C are the patient's identity and diagnosis, respectively; X is the information intended for disclosure; and Z is the sanitized (encoded) information that gets disclosed. 81
3.8 Demonstrating the "folding" intuition. (a) The data distribution, before sanitizing; and (b) the sanitized data distribution ("folded"). 85
3.9 The distribution of the raw (not sanitized) data, $p(x|c)$, per weight category, $c \in \mathbb{C}$. Note that the weight categories are not perfectly separable in $\mathbb{I}$. 90
3.10 The distribution of the sanitized data, $p(z|c)$, per weight category, $c \in \mathbb{C}$. Note that the different weight categories are now less distinguishable than before. 92

4.1 Visualization of the filter $\mathcal{F}(\cdot)$ with illustrative plots of p-values' probability density functions (PDFs) $f_1(\cdot)$ and $f_2(\cdot)$ demonstrating cases of no bias and bias, respectively. 108
4.2 Distribution of reported p-values in the exploration dataset. 115

List of Tables

1.1 Readmission rates for most frequent conditions and most frequent reason for rehospitalization for patients in Medicare Fee-for-service program; data for patients first discharged between October 1, 2003 and December, 31 2003 [Jencks et al., 2009]. . . 6

2.1 The demographics of the subjects. 26
2.2 Data collected by the CHF telemonitoring app. 27
2.3 The intervention message, depending on the estimated risk of clinical deterioration. 28
2.4 Taxonomy summary of the different health and fitness frameworks [Aranki et al., 2017b]. 34
2.5 The design objectives and principles in designing the Berkeley Telemonitoring framework. 39

3.1 The different weight categories defined by the corresponding Body Mass Index (BMI) percentiles within the same age and gender group for individuals of age 19 or less. This definition is consistent with the one provided by the CDC. 81
3.2 The confusion matrix of the classification of the raw (not sanitized) data. UW = Underweight, HW = Healthy Weight, OW = Overweight, OB = Obese. 90
3.3 The confusion matrix of the classification of the sanitized data. UW = Underweight, HW = Healthy Weight, OW = Overweight, OB = Obese. 92

4.1 Different scenarios of true number of trials and true number of successful ones, with the Type I error in each scenario and whether we reject H_0 with significance level 0.001. 101
4.2 The results using partitions with singleton blocks. Each row describes the results for a subset of the dataset corresponding to the set of DOF stated in the first column. 117

1 List of the nineteen journals surveyed from American Psychological Association (APA). Some journals published their first issue after the year 2002, in which case we mention in parentheses the year of the first issue published by that journal. The journals list is sorted according to the number of publications with experiments surveyed in our analysis. 142

List of Algorithms

2.1 A code snippet to create backables for heart-beat rate and cadence data, and registering them with a backup cabinet. . . . 40
2.2 A code snippet to create a backup cabinet located in the private app storage area. . 41
2.3 An example listener for heart-beat rate and blood pressure data. . . . 42
2.4 A standard snippet for communicating with ISO/IEEE 11073 Personal Health Device (PHD) enabled devices. . . . 42
2.5 A code snippet that uses a cadence estimator, requesting an update every 3 seconds. 44
2.6 A code snippet to render all survey nodes in a given survey. . . . 45
2.7 A code snippet to start a telemonitoring server that handles one job using the `MyJobListener` job listener. . . . 46
2.8 A code snippet to obtain a `ServerHandler` for the telemonitoring app to communicate with the server in a fault tolerant way; and an example of sending a data job. . . . 47
2.9 A code snippet that implements a job listener `MyJobListener` (see Algorithm 2.7) that stores EE data in a Structured Query Language (SQL) database, using a `EETableModifier`. . . . 48
3.1 calcMI $(R, D, p(c))$ – the objective function . . . 88
3.2 learnPMF $(\Theta, R_{\text{gen}}, D)$ – the learning procedure . . . 89
3.3 The MATLAB code for learning the Privacy Mapping Function (PMF) from the BMX_G data using the PDI toolbox. . . . 91
4.1 calcExpExp (θ, n, e, H_0) – Calculates Equation (4.6). . . . 111
4.2 calcProdMaxExpExp $(\theta, D, H_0, \mathcal{N}_p)$ – Calculates Equation (4.8) using a partition of N. . . . 113
4.3 ttestDatasetBound $(D, H_0, \mathcal{N}_p)$ – Calculates an upper bound of $p(D|H_0)$ . . . 113

Chapter 1

Introduction

An apple a day keeps the doctor away.

– cf. Welsh proverb, c. 1860

1.1 Genesis

In most regards, the nature of the healthcare model in the United States of America (US) is reactive. That is, we usually seek medical advice or care only *after* we perceive a deterioration in our health status. In other words, we first get sick, then go see a doctor. The general workflow of this healthcare model is depicted in Figure 1.2a. More generally, under the current healthcare model in the US, there is little encouragement for people to invest into their health before notable symptoms indicating a serious condition take place, with a few exceptions. Two of these exceptions concern two of the most common cancers in the US. Specifically, i) prostate cancer, which is the most common cancer in men in the US, aside from skin cancer; and ii) breast cancer, which is the most common cancer in women in the US [US Cancer Statistics Working Group et al., 2016].

According to the National Cancer Institute (NCI), many healthcare professionals encourage yearly prostate cancer screening for men starting from the age of 40 for high prostate-cancer risk groups (such as African American men) and 50 for others [National Cancer Insitite, 2012]. Prostate-specific antigen (PSA) screening is one of the two common preemptive prostate cancer screening procedures (the other being the digital rectal exam). Even though the accuracy of the PSA screening as a procedure for detecting prostate cancer has been in question since its introduction [Thompson et al., 2003], the literature provides some evidence to its efficacy. Welch and Albertsen estimated that as a result of introducing PSA screening in 1987, an additional 1.3 million men in the US were diagnosed with prostate cancer; with more than 1 million treated [Welch and Albertsen, 2009]. Moreover, there has been a reduction of about 33 % in the prostate cancer mortality rate in the US from 1992 to 2004 [Centers for Disease Control and Prevention, 1979-1998, 1999-2015; Jemal et al., 2008]. A survey by Etzioni et al. predicted that the plausible contribution of PSA screening to this decline is between 45 % and 70 % [Etzioni et al., 2008]. More recently, the prostate cancer mortality rates in the US have dropped by more than half in 2015, compared to 1992 [Centers for Disease Control and Prevention, 1979-1998, 1999-2015]. The mortality rates caused by prostate cancer in the US from 1968 until 2015 are depicted in Figure 1.1.

Similarly, The United States Preventive Services Task Force and NCI encourage women to regularly get screened for breast cancer after turning 50 years of age, even before any symptoms arise [National Cancer Insitite, 2017]. Women who are 40 to 49 years old are also encouraged to consult with their doctors regarding when to start and how often to get a breast cancer screening. Mammography is the most common screening modality for breast cancer, which in essence utilizes an X-ray image of the breast [Egan, 1962]. The use of mammography as a breast screening technique became widely adopted after Shapiro et al. demonstrated its impact on breast-cancer caused mortality in 1966 [Shapiro et al., 1966]. Shapiro et al. later reported the 10 to 14 year survival rates of breast cancer based on the same longitudinal study, with about 30 % lower 10-year mortality in subjects that underwent screening with mammography as opposed to those who didn't [Shapiro et al., 1982]. Since 1968, the breast cancer mortality rate dropped by about 37 % [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015]. Figure 1.1 also depicts the mortality rates caused by breast cancer in the US from 1968 until 2015.

To generalize these two examples, one can entertain the idea of a general proactive healthcare model. Under such a model, intervention is provided upon a reasonable expectation of a deterioration in the health status of a person. In other words, medical care is sought before symptoms arise,

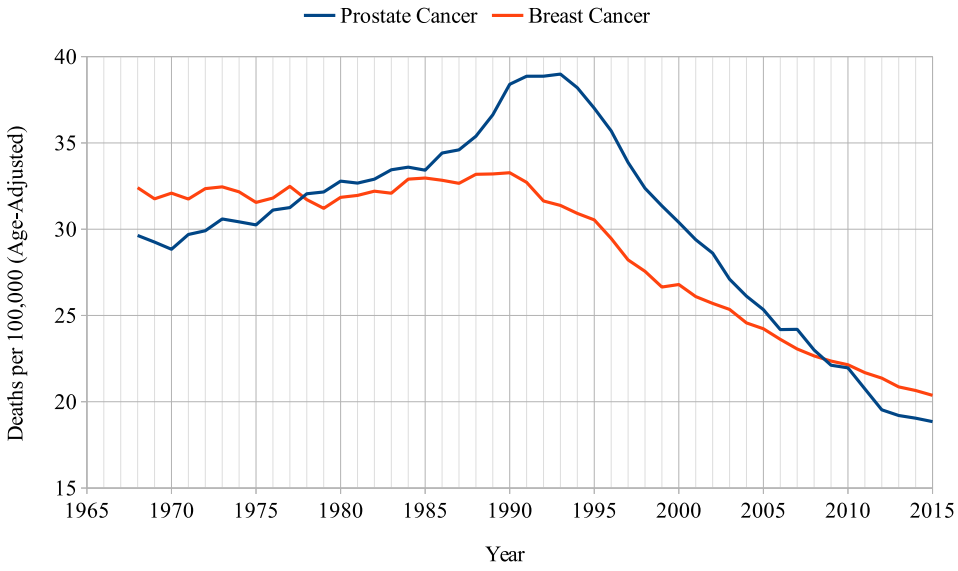


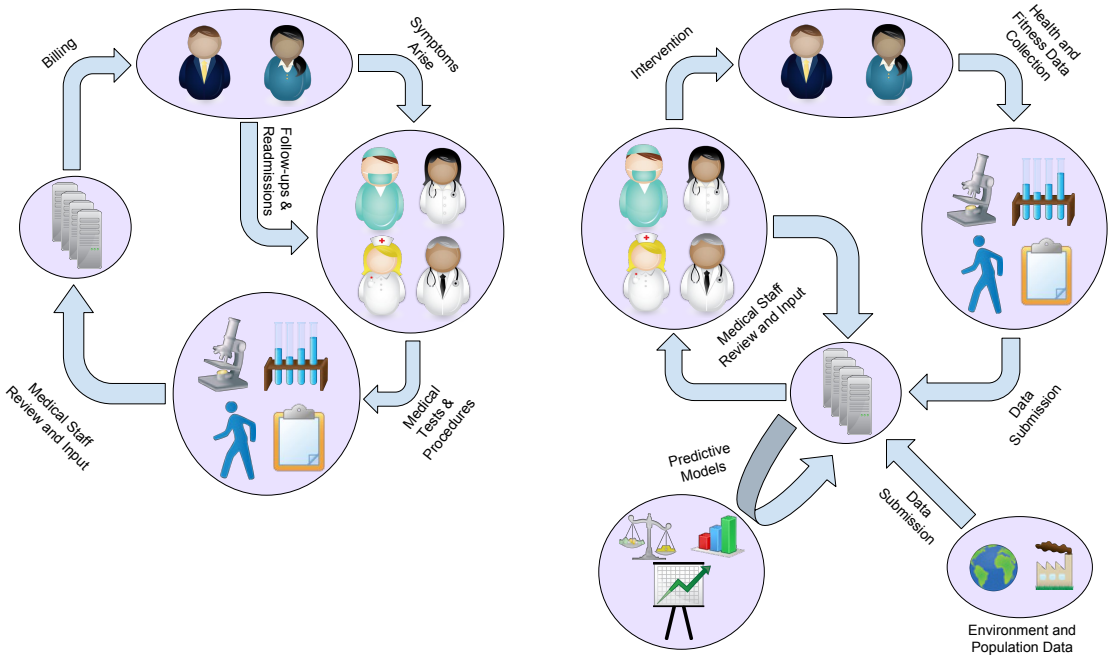
Figure 1.1: Age-adjusted mortality rates for prostate cancer (men) and breast cancer (women) between the years 1968 and 2015. The data were obtained from the Center for Disease Control and Prevention (CDC) [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015].

or before getting sick. Throughout this dissertation we will refer to this healthcare model as the *predictive healthcare model* or simply *predictive medicine*. As the name suggests, in the predictive healthcare model, predictive models are employed to forecast deteriorations in the health status of an individual. The predictive models can be calibrated and tailored to each person or a group of people with similar characteristics. If a deterioration is predicted, intervention can be provided to the person in question.

Other names for the described healthcare model exist in the literature (with possibly some nuanced differences), including *precision medicine*, *personalized medicine* and *preventative healthcare*. In a way, the prostate and breast cancer examples given above follow the preventative healthcare model. In general, the predictive healthcare model has the potential to i) improvement in individuals' well being; ii) reduction in healthcare costs; iii) reduction in readmission rates in chronic health conditions; and ultimately iv) reduction in all-cause mortality.

The generic workflow in the predictive healthcare model–depicted in Figure 1.2b–follows a feedback loop as follows. A patient's health data are submitted to a server that contains data about other individuals, data about the environment and clinical predictive models. The predictive models are used to assess the clinical risk of deterioration of the patient in question using the patient's data and the data about the environment. Medical personnel can then view the data and the clinical risk assessment results (from the predictive models) and provide further input. If the output of this cycle of clinical risk assessment deems that the patient is in high risk of clinical deterioration, a medical intervention is sought in an effort to circumvent this predicted deterioration. Simply put, in this model, medical attention is provided before the patient's health status worsens.

The predictive healthcare model is not yet realized as the standard of care. In this dissertation, we examine 3 outstanding challenges in the realization of an effective predictive healthcare model. Namely, these challenges are i) attaining accurate medical predictive models; ii) unbiased validation



(a) (b)

Figure 1.2: (a) The general workflow of the current healthcare model; and (b) the general workflow in the predictive healthcare model.

of said models; and iii) privacy in the predictive healthcare model. Before we delve into these challenges, we motivate the predictive healthcare model by looking at some of the current problems in the existing healthcare model in the US; problems that the predictive healthcare model can help address.

The rest of this chapter is organized as follows. We first examine rehospitalization rates in the US in Section 1.1.1, mortality rates in the US in Section 1.1.2 and the cost of healthcare in the US in Section 1.1.3. Afterwards, we introduce some of the challenges in realizing an effective predictive healthcare model in Section 1.2, laying the grounds for the rest of the dissertation. In Section 1.3, we list the original contributions of this work and we close this chapter by describing the organization of the rest of this dissertation in Section 1.4.

1.1.1 Readmission Rates

Unplanned reshopitalization carry an extra and often preventable cost to the healthcare system in the US [Keenan et al., 2008]. The most common discharge diagnosis for beneficiaries of Medicare who get rehospitalized (for any reason) is congestive heart failure (CHF) [Centers for Medicare & Medicaid Services, 2006].[1] Some of the possible factors for unplanned rehospitalization in the US include i) premature hospital discharge; ii) complications that manifest after discharge; iii) poor care transitions; and iv) underutilization of medical interventions [Keenan et al., 2008]. In this section, we look at the phenomenon of unplanned rehospitalization in the US in more detail.

We refer to unplanned rehospitalization after discharge by readmissions. Readmission rate is

[1]The Medicare program, which is funded by taxpayers, covered 46 million people of age 65 or older, and 9 million younger people with disabilities in the year 2015.

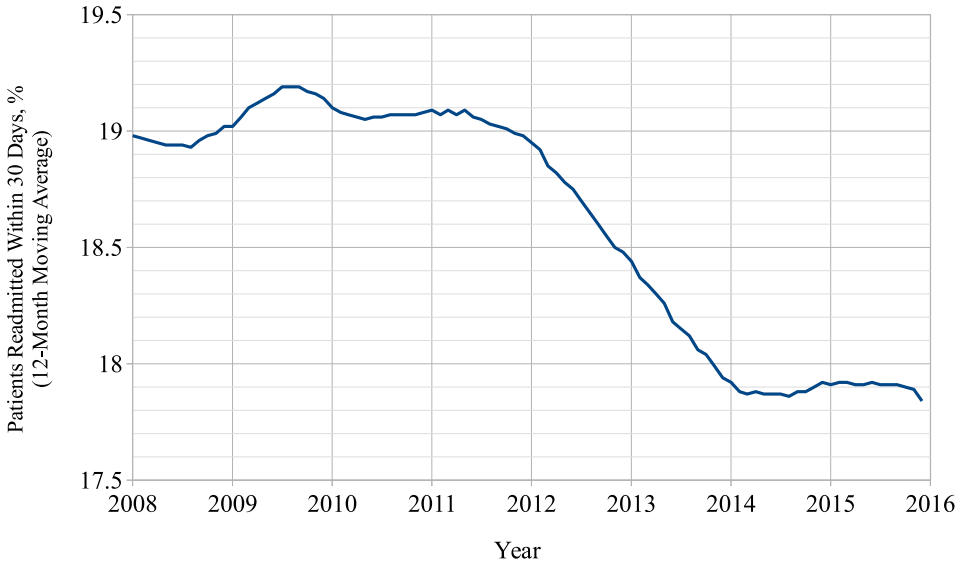


Figure 1.3: Medicare 30-Day, all-condition hospital readmission rate. Data were provided by the CMS in written communication to President Obama. The plotted series reflects a 12-month moving average of the hospital readmission rates reported for discharges occurring in each month [Obama, 2016].

defined as the percentage of patients who, within a pre-defined period of time of initial discharge, get rehospitalized without prior scheduling. Keenan et al. derived a model for hospital risk-standardized 30-day all-cause readmission rate for patients hospitalized with heart failure model [Keenan et al., 2008]. A generalized version of this model was later adopted by the Centers for Medicare & Medicaid Services (CMS) as the standard for measuring and reporting readmission data [Horwitz et al., 2011]. Moreover, the Patient Protection and Affordable Care Act (PPACA) passed under President Obama added the Readmissions Reduction Program, which also uses the risk-standardized readmission rate as the standard for measuring and reporting readmission data [US Congress, 2010].

To provide some concrete numbers for readmission rates, we focus on the older population in the US, who are covered by and large by the Medicare program. In Figure 1.3, we present the 12-month moving average of all-condition readmission rates, within a window of 30 days, for patients covered by Medicare [Obama, 2016]. The plot in Figure 1.3 presents a data point per month. Despite the improvement in the average all-condition readmission rates, within a window of 30 days, following the enactment of PPACA in 2010, we believe that there is still room for improvement.

According to a survey by Jencks et al., 26.9 % of patients in Medicare Fee-for-Service program with CHF were readmitted within 30 days of the initial discharge between October 1, 2003 and December 31, 2003; 24.6 % of patients with psychosis in Medicare, 22.6 % of patients with chronic obstructive pulmonary disease (COPD) in Medicare and 20.1 % of patients with pneumonia in Medicare were readmitted within 30 days of discharge in the same period [Jencks et al., 2009]. Table 1.1 depicts the readmission rates for the most frequent conditions and the most frequent reason for rehospitalization for patients in Medicare fee-for-service program in the same period [Jencks et al., 2009]. In another analysis, Giamouzis et al. report that approximately 50 % of patients with CHF are readmitted within 6 months of discharge [Butler and Kalogeropoulos, 2008],

Condition at Index Discharge	30-Day Rehospitalization Rate (%)	Proportion of All Rehospitalizations (%)	Most Frequent Reason for Rehospitalization (Percentage of Rehospitalization)
Meidcal			
All	21.0	77.6	Heart failure (8.6)
CHF	26.9	7.6	CHF (37.0)
Pneumonia	20.1	6.3	Pneumonia (29.1)
COPD	22.6	4.0	COPD (36.2)
Psychoses	24.6	3.5	Psychoses (67.3)
Gastrointestinal problems	19.2	3.1	Psychoses (21.1)
Surgical			
All	15.6	22.4	CHF (6.0)
Cardiac stent placement	14.5	1.6	Cardiac stent placement (19.7)
Major hip or knee surgery	9.9	1.5	Aftercare (10.3)
Other vascular surgery	23.9	1.4	Other vascular surgery (14.8)
Major bowel surgery	16.6	1.0	Gastrointestinal problems (15.9)
Other hip or femur surgery	17.9	0.8	Pneumonia (9.7)

Table 1.1: Readmission rates for most frequent conditions and most frequent reason for rehospitalization for patients in Medicare Fee-for-service program; data for patients first discharged between October 1, 2003 and December, 31 2003 [Jencks et al., 2009].

and 70 % of these rehospitalizations are associated with worsening of the previously diagnosed CHF [Gheorghiade et al., 2005; Giamouzis et al., 2011]. Although the readmission rates reported here are primarily for Medicare patients, we argue that there is room for improvement. We argue that the predictive healthcare model can reduce these readmission rates by enabling intervention before the need for rehospitalization in the cases of clinical deterioration [for example Aranki et al., 2016b, for CHF readmission rates].

1.1.2 Mortality Rates

The mortality rates measure can serve as yet another gauge for the quality of healthcare. In this section, we examine this measure in the US. The survey by Jencks et al. reports that 3.5 % of patients in Medicare fee-for-service program died within 30 days of initial discharge–without rehospitalization–between the dates October 1, 2003 and December 31, 2003. Among the same patients in the same time period, the death rates–without rehospitalization–within 60 days, 90 days, 180 days and 365 days were 4.5 %, 5.1 %, 6.0 % and 6.8 %, respectively [Jencks et al., 2009].

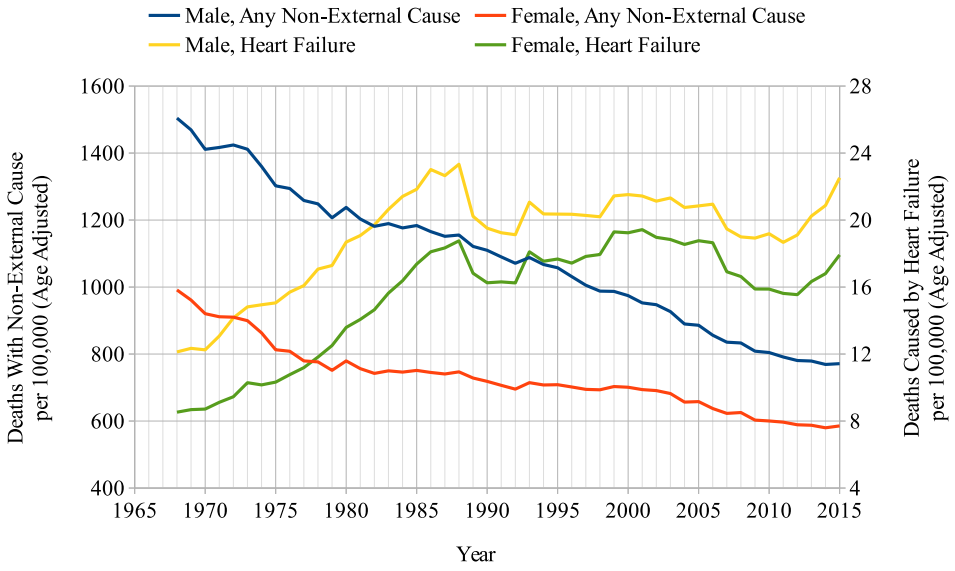


Figure 1.4: Age-adjusted mortality rates for i) deaths not caused by external causes; and ii) deaths caused by CHF in men and women between the years 1968 and 2015 in the US. The data were obtained from the CDC [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015].

Another way to look at mortality rates is to examine the number of deaths per 100,000 people. To get a general picture, we first look at deaths resulting from any cause that is not external, irrespective of hospitalization. External causes of death are reasons that are associated with events that occurred outside of the body. These causes include, for example, assault, self harm, accidents and legal intervention. Since 1968, these mortality rates in the US have dropped by about 49 % and 41 % for men and women, respectively. Even thought the non-external cause mortality rates have decreased since 1968, it is not a trend that is shared by all non-external causes, separately. For instance, since 1968, the CHF mortality rates in the US have increased by 86 % and 110 % for men and women, respectively [Centers for Disease Control and Prevention, 1968-1978, 1979-1998, 1999-2015]. Figure 1.4 depicts the age-adjusted mortality rates in the US for men and women, between the years 1968 and 2005, for deaths resulting from i) any cause that is not external; and ii) CHF. The example given above, for deaths caused by CHF, serves as evidence that there is some room for improvement in the healthcare system from the mortality rates perspective. We argue that the predictive healthcare model can reduce mortality rates in conditions like CHF by assessing risk of clinical deterioration and providing intervention in a timely fashion [Aranki et al., 2016b].

1.1.3 Cost of Healthcare

In addition to mortality rates and readmission rates, the cost of care in the US is high and increasing, and we argue that there is room for improvement in this category too. According to a report by CMS, the total spending on healthcare in the US was $3.2 trillion in 2015. That is, the average healthcare cost to a person living in the US was $9,990 in 2015 [Centers for Medicare & Medicaid Services, 2015]. This constitutes an increase of 5.8 % compared to the $3 trillion total spending on healthcare in the preceeding year. These numbers amount to 17.8 % and 17.4 % of

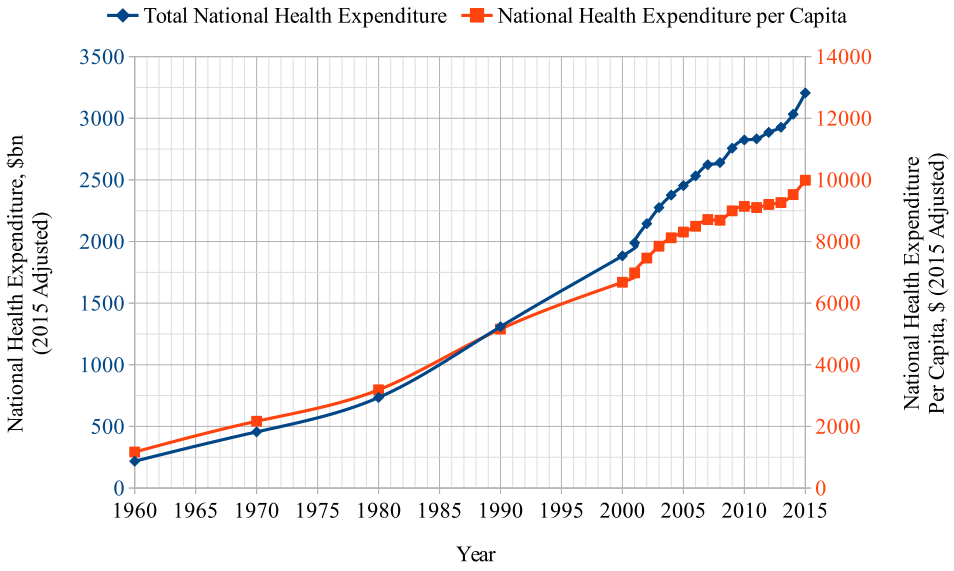


Figure 1.5: The national health expenditure in the US (total and per capita) since 1960, adjusted to 2015 USD.

the overall US gross domestic product (GDP), respectively. For example, Jencks et al. estimated that financial burden of readmissions on Medicare is $17 billion annually [Jencks et al., 2009]. Furthermore, Commonwealth Fund estimated in 2006 that Medicare could save $1.9 billion annually, if national readmission rates were brought down to the levels achieved by the top-performing regions [Commonwealth Fund, 2006]. As late as 2011, Commonwealth Fund reported that based on an analysis by the Medicare Payment Advisory Commission, Medicare can save $12 billion by reducing hospital readmissions, with additional savings possible from reductions in hospitalizations among the under-65 population [Commonwealth Fund, 2011]. More broadly, Figure 1.5 depicts i) the total national health expenditure in the US; and ii) the average healthcare expenditure to a person living in the US, between the years 1960 and 2015 (adjusted to 2015 United States Dollar (USD)).

In 2015, the national health expenditure in the US was almost 15 times that of 1965–adjusted to inflation. In contrast, in 2015, the national health expenditure per capita was about 8.5 times that of 1965–adjusted to inflation. To give perspective to the magnitude of the healthcare spending, compared to the 17.8 % in 2015, the national health expenditures in 1965 amounted to 5 % of the US GDP in that year. The plot of the proportion of the US national health expenditure from the GDP, since 1960, is depicted in Figure 1.6.

The cost of healthcare in the US has a lot of room for improvement. We believe that the predictive healthcare model can significantly lower the cost of healthcare in the US. Particularly, this can be achieved by reducing readmission rates through early prediction of clinical deterioration and subsequent medical intervention as discussed earlier.

1.2 The Gaps Towards Predictive Medicine

The predictive healthcare model is not yet fully realized. The burning question is *what is standing in the way of fully realizing an effective predictive healthcare model?* This question is the core question that we try to address in this dissertation. Unfortunately, as it turns out, the full potential

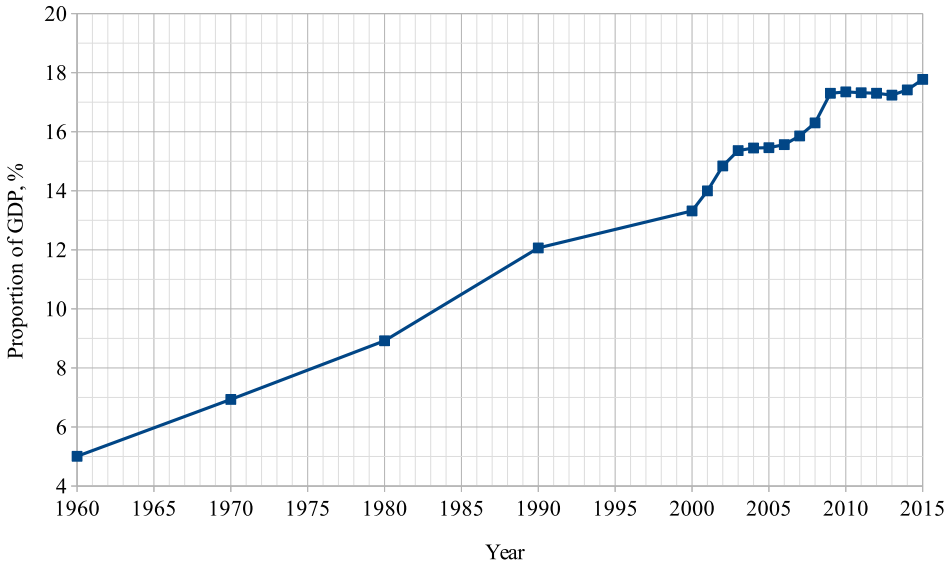


Figure 1.6: The proportion of the US national health expenditure from the GDP, since 1960.

of this question is larger than any single dissertation can address. The reason is that there are many outstanding challenges in achieving this model in the US. Particularly, these challenges include ones in the i) regulation; ii) scientific; iii) engineering; and iv) social realms. As a result, we will focus on only some challenges. In this section, we describe the challenges that we will address in this dissertation.

In general, achieving improvements in clinical and financial outcomes in the healthcare domain can take place on multiple fronts. For starters, the role of regulation in tackling these challenges is important. For instance, in an effort to reduce the rising cost of healthcare in the US and improve healthcare coverage, the US legislature passed PPACA in March 2010 under President Barack Obama. For instance, in order to achieve reduction in cost, one of the main mechanisms was to reduce the rates of readmission in chronic health conditions [US Congress, 2010; Obama, 2016].

Complementary to regulation, the role of science in improving clinical outcomes is primary. We argue that a predictive healthcare model can help realize better improvements to i) cost of care; ii) readmission rates; iii) mortality rates; and iv) overall quality of care. First, a key component in the predictive healthcare is the ability to make predictions about the clinical status of a patient [Aranki et al., 2016b]. As a result, there is a need to obtain accurate clinical predictive models, preferably tuned and calibrated to each patient or group of patients separately (see Figure 1.2b). Devising and validating these predictive models often require longitudinal studies that are expensive and challenging to design. The classical example of such a study is the Framingham Heart Study, which aims to unveil risk factors and the epidemiology of cardiovascular diseases. The study began in 1948 and, among many findings, revealed the effects of diet, exercise and aspirin on heart disease. Moreover, the study resulted in a score that estimates the 10-year cardiovascular risk of individuals, including those without known cardiovascular disease; which in our terms is a predictive model [D'Agostino et al., 2008]. The Framingham Heart Study serves as an example for i) the need for long-term longitudinal studies in order to devise predictive models and clinically validate them; ii) the complexity of such long-term longitudinal studies (in terms of cost, administration, etc.); and iii) the potential side benefits of such studies (e.g., revealing effect of diet on heart disease).

There are two factors in epidemiological studies that make them harder to implement. Concretely, these studies require observing health-related data and clinical outcomes i) of a *large sample of people*, ii) for a *long period of time*. In order to achieve a *ubiquitous* predictive healthcare model, the cost of research and validation is going to be high. Alternatively, we can rethink the process of monitoring large number of subjects for long periods of time for health-related data and clinical outcomes. Therefore, the first challenge that we address in this dissertation is

> Devising a streamlined and affordable process for data collection and monitoring of subjects applicable for epidemiological studies.

An extended abstract of this part of the dissertation is presented in Section 1.2.1.

Second, in order to achieve a *reliable* predictive healthcare model, we have to validate the devised predictive models and the scientific findings. For starters, it is vital that any results that are deemed scientifically valid be *reproducible*. Complementary to reproducibility, in order to draw the correct conclusions about scientific hypotheses, it is vital to see an *unbiased* image of all studies and experiments attempting to validate a hypothesis. In particular, it is as important to publish studies that did not yield a statistically significant effect size (studies where the p-value is higher than the significance level) as it is to publish studies yielding statistically significant effect sizes. There is a longstanding belief in the scientific community that the scientific literature is biased against publications not yielding statistically significant effect sizes. Therefore, the third challenge that we address in this dissertation is

> Devising a meta-validation process for scientific literature that tests for publication bias.

An extended abstract of this part of the dissertation is presented in Section 1.2.2.

Lastly, in order to achieve a *prevalent* predictive healthcare model, we have to consider consumer privacy. Ensuring that consumer privacy is protected will, among other things, allow researchers to reach a larger sample of people in epidemiological studies. Furthermore, we argue that ensuring consumer privacy will result in more *unbiased* sample of subjects and a more unbiased sample of data for epidemiological studies [Warner, 1965]. Therefore, the second challenge that we address in this dissertation is

> What privacy models are applicable for the predictive healthcare model depicted in Figure 1.2b? Particularly, what privacy models are applicable for the devised process of data collection and monitoring in the first challenge?

An extended abstract of this part of the dissertation is presented in Section 1.2.3.

1.2.1 The Need for Reliable Data

In order to make a feasible predictive healthcare model, there is a need for accurate models capable of predicting clinical deterioration. In order to develop those predictive models, there is a need of a process to measure human physical variables in an accurate, unobtrusive, non-invasive, continuous and cost-effective manner. The process needs to be designed in a way that is compatible with longitudinal epidemiological studies, allowing a large number of subjects to be monitored for long periods of time.

We are seeing an influx of consumer fitness, health and medical devices that enable the collection of human physical variables including activity levels, blood pressure, heart rate and others. In

the presence of such measurement devices, consumers can ultimately be remotely monitored by healthcare providers and decisions on intervention can be derived from predictions made using these measurements. Unfortunately, many of these devices are currently inaccurate for medical purposes. Furthermore, developing research systems that utilize such sensors in order to collect and ultimately predict health status is not an easy task as it requires knowledge in both the medical sciences and in engineering. We address the challenge of creating a research-oriented remote health monitoring framework in Chapter 2.

In this part of the dissertation, we examine the notion of telemonitoring as means of collecting health-related data about subjects in longitudinal epidemiological studies. Telemonitoring can be further used as a system for risk-assessment of clinical deterioration, once predictive models are devised. This is because the general workflow of telemonitoring systems follows the same workflow of the predictive healthcare model presented in Figure 1.2b, allowing it to fit nicely in this healthcare model. The advantage of telemonitoring is that it is relatively cheap to deploy to a large sample of subjects in studies [McConnell et al., 2017, for example]. The challenge with telemonitoring, on the other hand, is to design the system to collect *reliable* data for medical purposes. We present a feasibility study that we conducted in collaboration with Northwestern Medical Faculty Foundation and New York University in which we remotely monitored 15 patients with CHF using a custom smartphone application [Aranki et al., 2014, 2016b].

We draw the lessons learned from that study and argue the need for systematic treatement of these issues, giving birth to the Berkeley Telemonitoring project [Aranki et al., 2016a].[2] We first describe the design objectives of a general-purpose, research-oriented framework for mobile-based remote monitoring. We then describe the resulting modular design of the Berkeley Telemonitoring framework and some of its implementation details. Finally, we provide a taxonomy of the existing solutions for remote monitoring, including the Berkeley Telemonitoring framework.

As empirical evidence of the usability and practicality of the Berkeley Telemonitoring framework, we present the *RunningCoach* app that was built using the Berkeley Telemonitoring framework. RunningCoach aims at candence-oriented training for long-distance runners [Aranki et al., 2017b]. We close this part of the dissertation by describing the findings from the field study conducted using the RunningCoach app.

1.2.2 The Need for Reliable Science

Validating the predictive power of any model requires empirical studies. The main purpose of these studies is to measure the accuracy of these models. Primarily, these studies consist of a limited number of subjects (compared to the whole population) and attempt to derive estimates of the model's accuracy on the whole population by observations and measurements made on the participating subjects. The backbone of these extrapolations is the assumption of reproducibility. That is, if the study were to be run by other groups of independent researchers on different groups of subjects, similar findings would be found up to a probability of error.

In addition to reproducibility–and complementary to it, in order to draw a decisive conclusion on the efficacy of any treatement or the predictive power of any model, it is vital that *all successful and unsuccessful trials* concerning the same hypothesis be reported. Otherwise, the drawn conclusions are merely anectodal due to inadvertent cherry-picking of results. The case where some of the unsuccessful attempts are not reported (more so than those which were successful) is often dubbed as *publication bias*. Unfortunately, there is a strong belief in the scientific community that

[2]The Berkeley Telemonitoring project: `https://telemonitoring.berkeley.edu`

publication bias exists in the literature of empirical research. Therefore, if this problem indeed exists, it hinders scientific progress in many fields, including the advancement of predictive medicine. We address the identification of publication bias and discuss efforts to limit it in Chapter 4.

In this part of the dissertation, we address the problem of detecting publication bias in the literature of empirical and experimental sciences. Publication bias is the phenomenon where the scientific literature contains publications showing statistically significant results more than their actual proportions in reality. In another view, studies showing statistically significant results are more likely to be published than studies with statistically insignificant, or even null results.

To illustrate this problem, consider the following example. If one were to test the (false) hypothesis that Daniel Aranki has psychic abilities in that he could tell the color of any playing card merely by looking at its back, one can design the following protocol. Draw N playing cards at random with replacement, and each time have Daniel guess the color of the card. There is a $\frac{1}{2^N} > 0$ chance that Daniel can guess the colors of all N cards under the null hypothesis that Daniel doesn't possess psychic powers. Now let us say that 2^N different people ran that study, independently, and only one succeeded to show that Daniel guessed all N card colors correctly (The expected number of such studies is 1). If all 2^N studies reported their findings, we are most likely going to dispute that Daniel has psychic powers (for large N). However, if only successful trials were to report their findings (in this case, the lone successful study), we would have a very limited and biased view on the truth of the matter; and are more likely to consider the possibility that Daniel has psychic powers. The latter case depicts an extreme case of publication bias.

In this dissertation, we describe the problem in a quantitative way. We use this description to define a measure on datasets under the null hypothesis that no publication bias exists. This measure in turn leads to a statistical test that yields an upper bound on the probability of observing the data set under the null hypothesis.

The derived theory is verified empirically using a dataset of 3,721 publications (with 23,117 hypothesis tests) from 12 journals of the American Psychologists Association between the years 2002 and 2012. We find that the probability of observing that dataset under the hypothesis of absence of publication bias is lower than $\frac{1}{500}$.

1.2.3 Privacy in Predictive Healthcare

Consumers are voicing increasing concerns regarding their health-related privacy, especially with the rise of electronic health records and health information technology in general [Bishop et al., 2005; Hsiao and Hing, 2012; Hussain et al., 2015]. Therefore, it is vital to address privacy in the new age of technologies for predictive medicine; particularly because a) it is hard to develop and test predictive models without wide and unbiased adoption of these technologies in the research stages; b) such technologies collect health and behavioral data in a continuous manner; and c) such technologies disclose said data with healthcare providers. In Chapter 3, we address the issue of preserving consumer privacy in predictive medicine in general and in technologies of mobile health (mHealth) telemonitoring in particular.

In this part of the dissertation, we claim that privacy by design is essential for consumer protection and wider adoption of such technology. We show preliminary evidence that consumers trust health technology researchers with their health information at similar levels to their healthcare providers, and at a significantly higher level than insurance companies, for example [Aranki et al., 2016b].

We then present the view of privacy under which data about an individual may be used to infer other undisclosed pieces of information about the same individual or even others. For example, in

a telemonitoring scenario, a patient may be disclosing continuous respitory rate data. If such data may be later used to infer whether the patient is a smoker or not, then the patient may consider that to be intrusive from a privacy point of view (assuming the patient considers smoking status to be a private piece of information). We claim that this is true regardless of whether the patient considers respitory rate information to be inherently private on its own right. Under this view of privacy, where unauthorized *interpretation of data* needs to be prevented, we present the framework for Private Disclosure of Information (PDI), a semantic privacy model that is applicable in mHealth telemonitoring settings. PDI aims at limiting the ability of a third-party to infer sensitive pieces of information about the patient while still allowing the healthcare provider to interpret and use the data in an accurate way [Aranki and Bajcsy, 2015]. We derive the theory of PDI, prove some of its properties and demonstrate its effectiveness empirically on health data obtained from the CDC.

1.3 Main Contributions

The research presented herein provides several contributions in the fields of evidence-based science, precision medicine, and privacy. Concretely, there are three main contributions of the research presented in this dissertation. First, we introduce a general-purpose framework for mHealth telemonitoring, designed in order to facilitate and streamline the process of remote monitoring of patients after their discharge. This framework is part of the Berkeley Telemonitoring project [Aranki et al., 2016a, 2017a]. The framework, however, may be used to collect data for longitudinal epidemiological studies, streamlining their design and roll-out. A more detailed discussion of the framework is presented in Chapter 2.

The second main contribution of this work is a privacy framework that is applicable to telemonitoring systems, named PDI. The PDI framework treats the communicated data as private information not because they are inherently secret, but because that data can be used to infer some private information about the subjects, that is not communicated through the telemonitoring system. We present the framework, prove some of its properties and present a general-purpose MATLAB toolbox that can be used to train the PDI framework from data. A more detailed discussion of the PDI framework is presented in Chapter 3.

The third main contribution of this work is a statistical test for detecting publication bias in scientific publications employing the Student t-test. The method defines a bias test statistic for a set of observed publications, based on their effect sizes and Degrees of Freedom (DOFs). Given a bias test statistic b, the method yields an upper bound on the probability of observing a dataset with a bias test statistic at least as extreme as b, under the null hypothesis that no publication bias exists. We further present a MATLAB implementation of the test. A more detailed discussion of this statistical test is presented in Chapter 4.

1.4 Dissertation Organization

The rest of this dissertation is organized as follows. We start by addressing the challenge of obtaining reliable health data in a streamlined fashion in Chapter 2, and introduce the Berkeley Telemonitoring framework. In Chapter 3, we address the issue of consumer privacy in predictive medicine and in mHealth telemonitoring; and introduce the PDI framework. We then address the question of detecting publication bias in the literature of empirical research in Chapter 4. We close

the dissertation in Chapter 5 by presenting some final thoughts on predictive medicine, privacy and bias in the scientific literature.

The dissertation is written in a manner that allows the reader to read each chapter separately, immediately after the introduction. That is, all the chapters in this book are self-contained and require reading only Chapter 1. The only possible exception to this is Chapter 5, which is still self-contained but benefits from reading the rest of the dissertation for context.

Bibliography

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D., Kurillo, G., and Bajcsy, R. Smartphone Based Real-Time Health Monitoring and Intervention. In Khan, S. U., Zomaya, A. Y., and Abbas, A. (eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, chap. ?? Springer, 2017a. In press.

Aranki, D., Kurillo, G., Mani, A., Azar, P., van Gaalen, J., Peng, Q., Nigam, P., Reddy, M. P., Sankavaram, S., Wu, Q., and Bajcsy, R. A telemonitoring framework for android devices. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 282–291. IEEE, 2016a.

Aranki, D., Kurillo, G., Sarver, H., Song, E., Asuncion, C., Serven, L., Balakrishnan, U., and Bajcsy, R. RunningCoach – Cadence-Oriented Training Application for Long-Distance Runners, 2017b. In preperation.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Continuous, real-time, tele-monitoring of patients with chronic heart-failure: lessons learned from a pilot study. In *Proceedings of the 9th International Conference on Body Area Networks*, pp. 135–141. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-Time Tele-Monitoring of Patients with Chronic Heart-Failure Using a Smartphone: Lessons Learned. *IEEE Transactions on Affective Computing*, vol. 7(3):pp. 206–219, July 2016b. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2554118.

Bishop, L., Holmes, B. J., and Kelley, C. M. National consumer health privacy survey 2005. *California HealthCare Foundation, Oakland, CA*, 2005.

Butler, J. and Kalogeropoulos, A. Worsening heart failure hospitalization epidemic. *Journal of the American College of Cardiology*, vol. 52(6):pp. 435–437, 2008.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1968-1978. 1968-1978. CDC WONDER Online Database, compiled from Compressed Mortality File CMF 1968-1988, Series 20, No. 2A, 2000. Accessed: Apr 3, 2017, URL `http://wonder.cdc.gov/cmf-icd8.html`.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1979-1998. 1979-1998. CDC WONDER Online Database, compiled from Compressed Mortality File CMF 1968-1988, Series 20, No. 2A, 2000 and CMF 1989-1998, Series 20, No. 2E, 2003. Accessed: Apr 3, 2017, URL `http://wonder.cdc.gov/cmf-icd9.html`.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1999-2015. December 1999-2015. CDC WONDER Online Database, compiled from Compressed Mortality File 1999-2015 Series 20 No. 2U, 2016, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed: Apr 3, 2017, URL `http://wonder.cdc.gov/cmf-icd10.html`.

Centers for Medicare & Medicaid Services. Medicare Ranking for all Short-Stay Hospitals by Discharges Fiscal Year 2005 versus 2004. September 2006. Accessed: Apr 3, 2017, URL `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/SSDischarges0405.pdf`.

Centers for Medicare & Medicaid Services. National health expenditures. 2015. Accessed: Mar 2017, URL `https://www.cms.gov/Research-Statistics-Data-and-systems/Statistics-Trends-and-reports/NationalHealthExpendData/Downloads/highlights.pdf`.

Commonwealth Fund. Why not the best? Results from the national scorecard on US health system performance, 2006. *New York: The Commonwealth Fund*, September 2006. Accessed: Apr 3, 2017, URL `http://www.commonwealthfund.org/publications/fund-reports/2006/sep/why-not-the-best--results-from-a-national-scorecard-on-u-s--health-system-performance`.

Commonwealth Fund. Why not the best? Results from the national scorecard on US health system performance, 2011. *New York: The Commonwealth Fund*, October 2011. Accessed: Apr 3, 2017, URL `http://www.commonwealthfund.org/publications/fund-reports/2011/oct/why-not-the-best-2011`.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, vol. 117(6):pp. 743–753, 2008. ISSN 0009-7322. doi:10.1161/CIRCULATIONAHA.107.699579. URL `http://circ.ahajournals.org/content/117/6/743`.

Egan, R. L. Mammography, an aid to diagnosis of breast carcinoma. *JAMA*, vol. 182(8):pp. 839–843, 1962.

Etzioni, R., Tsodikov, A., Mariotto, A., Szabo, A., Falcon, S., Wegelin, J., Karnofski, K., Gulati, R., Penson, D. F., Feuer, E., et al. Quantifying the role of PSA screening in the US prostate cancer mortality decline. *Cancer Causes & Control*, vol. 19(2):pp. 175–181, 2008.

Gheorghiade, M., Zannad, F., Sopko, G., Klein, L., Piña, I. L., Konstam, M. A., Massie, B. M., Roland, E., Targum, S., Collins, S. P., et al. Acute heart failure syndromes. *Circulation*, vol. 112(25):pp. 3958–3968, 2005.

Giamouzis, G., Kalogeropoulos, A., Georgiopoulou, V. V., Laskar, S., Smith, A. L., Dunbar, S. B., Triposkiadis, F., and Butler, J. Hospitalization epidemic in patients with heart failure: risk factors, risk prediction, knowledge gaps, and future directions. *Journal of Cardiac Failure*, vol. 17(1):pp. 54–75, 2011.

Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., Conover, M., Montague, J., Dillaway, C., Bartczak, K., Ross, J., et al. Hospital-wide (all-condition) 30-day risk-standardized readmission measure. *Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Retrieved September*, vol. 10:p. 2012, 2011.

Hsiao, C.-J. and Hing, E. *Use and characteristics of electronic health record systems among office-based physician practices, United States, 2001-2012*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.

Hussain, M., Al-Haiqi, A., Zaidan, A., Zaidan, B., Kiah, M., Anuar, N. B., and Abdulnabi, M. The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations. *Computer Methods and Programs in Biomedicine*, vol. 122(3):pp. 393–408, 2015.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., and Thun, M. J. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, vol. 58(2):pp. 71–96, 2008.

Jencks, S. F., Williams, M. V., and Coleman, E. A. Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, vol. 360(14):pp. 1418–1428, 2009.

Keenan, P. S., Normand, S.-L. T., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., Schuur, J. D., Stauffer, B. D., Bernheim, S. M., Epstein, A. J., et al. An Administrative Claims Measure Suitable for Profiling Hospital Performance on the Basis of 30-Day All-Cause Readmission Rates Among Patients With Heart Failure. *Circulation: Cardiovascular Quality and Outcomes*, vol. 1(1):pp. 29–37, 2008.

McConnell, M. V., Shcherbina, A., Pavlovic, A., Homburger, J. R., Goldfeder, R. L., Waggot, D., Cho, M. K., Rosenberger, M. E., Haskell, W. L., Myers, J., et al. Feasibility of Obtaining Measures of Lifestyle From a Smartphone App: The MyHeart Counts Cardiovascular Health Study. *Jama cardiology*, vol. 2(1):pp. 67–76, 2017.

National Cancer Insitite. Prostate-Specific Antigen (PSA) Test. 2012. URL `https://www.cancer.gov/types/prostate/psa-fact-sheet`.

National Cancer Insitite. Breast Cancer Screening. 2017. URL `https://www.cancer.gov/types/breast/hp/breast-screening-pdq`.

Obama, B. United states health care reform: Progress to date and next steps. *JAMA*, vol. 316(5):pp. 525–532, 2016. doi:10.1001/jama.2016.9797. URL `http://dx.doi.org/10.1001/jama.2016.9797`.

Shapiro, S., Strax, P., and Venet, L. Evaluation of periodic breast cancer screening with mammography: methodology and early observations. *JAMA*, vol. 195(9):pp. 731–738, 1966.

Shapiro, S., Venet, W., Strax, P., Venet, L., and Roeser, R. Ten-to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*, vol. 69(2):pp. 349–355, 1982.

Thompson, I. M., Goodman, P. J., Tangen, C. M., Lucia, M. S., Miller, G. J., Ford, L. G., Lieber, M. M., Cespedes, R. D., Atkins, J. N., Lippman, S. M., et al. The influence of finasteride on the development of prostate cancer. *New England Journal of Medicine*, vol. 349(3):pp. 215–224, 2003.

US Cancer Statistics Working Group et al. United States Cancer Statistics: 1999–2013 Incidence and Mortality Web-based Report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, 2016. URL `https://nccd.cdc.gov/uscs/`.

US Congress. Patient Protection and Affordable Care Act. *Public Law*, (111-148), 2010.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60(309):pp. 63–69, 1965.

Welch, H. G. and Albertsen, P. C. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *Journal of the National Cancer Institute*, vol. 101(19):pp. 1325–1329, 2009.

Chapter 2

Telemonitoring for Predctive Medicine

> *The doctor of the future will give no medicine, but will instruct his patient in the care of the human frame, in diet and in the cause and prevention of disease.*
>
> – Thomas Edison, 1903

2.1 Introduction

One of the core building blocks of predictive medicine is accurate predictive models for diseases and health deterioration. In order to construct such predictive models, we need to understand the risk factors and causes of disease. The field of science that is concerned with the discovery, study and analysis of these causes and risk factors is called *epidemiology*. By nature, most epidemiological studies are longitudinal studies that span over a long period of time. As such, a major roadblock in achieving a predictive healthcare model is the prohibitive cost of epidemiological studies.

Some of the other building blocks of the predictive healthcare model, as depicted in Figure 2.1, are i) health-related data collection; ii) submission of such data to data-analysis data centers; iii) the analysis of such data, including assessment of risks by employing predictive models and input from healthcare professionals; and iv) medical intervention based on these predictions.

In light of this context, we address the following question in this chapter.

> *How can technology alleviate the prohibitive cost of epidemiological studies without compromising the reliability and integrity of these studies and the data collected therein?*

The key observation that we make is that a technology that implements the predictive healthcare model from Figure 2.1, apart from the prediction loop, can be used to streamline and standardize epidemiological studies. In addition, the same technology can serve as a building block for a complete system for predictive medicine. This can be done by complementing it with the predictive models that are extracted from the epidemiological studies. In essence, a technology that enables the collection and submission of health-related data in epidemiological studies, can later be used in the deployment of predictive medicine.

To be concrete, we consider a ubiquitous system that implements the highlighted parts of the predictive healthcare model depicted in Figure 2.1. In particular, this system enables i) the collection and submission of health-related data; ii) allowing healthcare professionals to view the data and provide input; and iii) delivering some forms of medical intervention. Throughout this dissertation, we call such a system a *telemonitoring system*. Once predictive models are developed, based on epidemiological findings, they can be incorporated in telemonitoring systems. Therefore, telemonitoring can be used to both i) streamline and standardize epidemiological studies in a cost-effective way; and ii) implement a full predictive healthcare model, once predictive models are incorporated. In this chapter, we study the design, implementation and efficacy of telemonitoring systems in general, and mobile health (mHealth) in particular.

The rest of this chapter is organized as follows. We start by investigating the feasibility, acceptability and requirements of telemonitoring. In our efforts to better understand those, we conducted a study of telemonitoring patients with congestive heart failure (CHF) in collaboration with Northwestern Medical Faculty Foundation (NMFF), Chicago, IL and New York University (NYU), New York, NY. The study, its findings and the lessons learned from it are described in Section 2.4. We then survey the literature for related work in the field of telemonitoring in general, and mHealth in particular in Section 2.2. The literature survey is followed by a survey

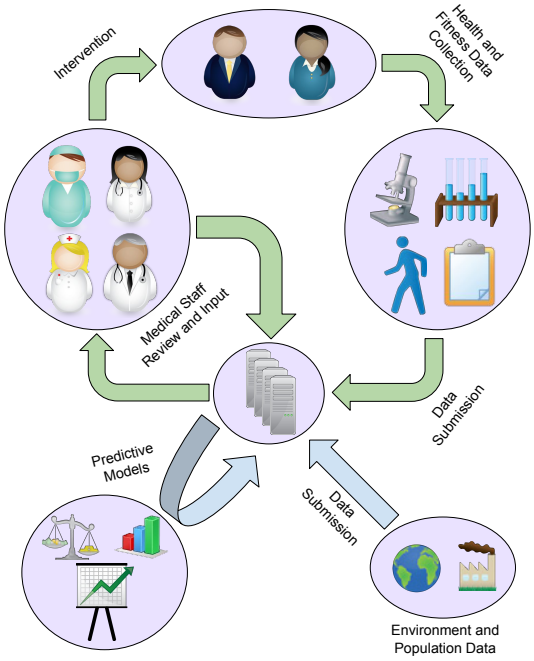


Figure 2.1: The general workflow in the predictive healthcare model. This model also coincides with the workflow of telemonitoring systems.

of commercial solutions, which attempt to address and implement telemonitoring and mHealth in Section 2.5. Afterwards, we briefly visit the concept of medical intervention in Section 2.3. We then summarize the challenges in the fields of mHealth and telemonitoring in Section 2.6. Subsequently, we present the Berkeley Telemonitoring project and the Berkeley Telemonitoring framework in Section 2.7 followed by a second study in Section 2.8, RunningCoach, which tested the usability of the Berkeley Telemonitoring framework and served as a proof-of-concept for it in fitness applications. Finally, we summarize the chapter and discuss its key points in Section 2.9.

The contributions of this chapter are i) providing evidence for proof-of-concept for the feasibility of telemonitoring in health and fitness applications with two studies (Sections 2.4 and 2.8); and ii) introducing the Berkeley Telemonitoring project and the Berkeley Telemonitoring framework (Section 2.7).

2.2 Feasibility and Related Work

In this section, we briefly survey the literature in three topics relevant to our discussion in this chapter: (a) clinical telemonitoring in patients with CHF; (b) use of smartphone apps in mHealth; and (c) measurement of physical activity. This survey provides us the context, motivation and background information for the Berkeley Telemonitoring framework, the telemonitoring solution presented in this chapter.

2.2.1 Telemonitoring in Patients with CHF

Since the predictive healthcare model is not quite achieved, we instead examine several examples of telemonitoring for proactive healthcare. In proactive healthcare, patients are encouraged to take proactive actions to reduce risks based on their health data, without necessarily having accurate

predictive models in place. These studies and systems gave rise to mHealth-based telemonitoring, which we examine in great detail in this chapter. In particular, we focus on CHF as a significant example for the benefits of telemonitoring and medical intervention, including reduction of cost of care and reduction of readmission rates. It is an impossible job to give justice to all the studies on telemonitoring while remaining brief. Therefore, we elected to summarize the findings of several systematic review studies in the field. We refer interested readers to the cited papers, and the references therein, for more details.

Telemonitoring in general, and for CHF purposes in particular, is older the inception of mHealth as a field. For example, In a randomized clinical trial that spanned from 2006 to 2009, Chaudhry et al. [2007] accomplished telemonitoring using a voice-response system that called patients with CHF, daily, to inquire about their symptoms and weight. There were 826 patients in the treatment group (to which telemonitoring was provided) and 827 patients in the control group (which received the usual care). The treatment group patients' physicians then proceeded to review the collected data to assess the risk of clinical deterioration and subsequently decide on medical intervention. Chaudhry et al. [2010] reported no significant differences in readmission or mortality rates within 180 days of enrollment, between the two groups.

As stated earlier, in the interest of brevity, we will focus mostly on systematic reviews of randomized clinical trials of telemonitoring in patients with CHF. Clark et al. [2007] published a systematic review of 14 of such randomized controlled trials. In these trials, 4262 total patients with CHF participated. The review reported an average reduction in CHF-caused readmission by 21% and an average reduction in all-cause mortality by 20%. The 95% confidence intervals were $11\% - 31\%$ and $8\% - 31\%$, respectively. It is noteworthy to mention that only 1 study collected information about daily physical activity, self-reported through means of telephone contact by a nurse. Moreover, only five of the reviewed studies collected daily information about vital signs (such as blood pressure, heart-beat rate and/or periodic electrocardiogram, or weight), for of which collected symptoms-related information as well (such as fatigue or shortness of breath).

In another review, Inglis [2010] reported on the findings of 30 studies of telemonitoring in patients with CHF. Similar to the trends observed from the previous reviews, only two studies contained a notion of activity monitoring in their protocol. In one of them, the activity monitoring of subjects was achieved through provided activity monitors that were only used for self-monitoring by subjects. That is, the activity data were never submitted to the medical team for analysis or intervention-decision-making [Galbreath et al., 2004]. The other study, a structured telephone support study that was included in the systematic review by Clark et al. [2007], only collected self-reported daily physical activity about the participating patients.

In yet another review, Giamouzis et al. [2012] analyzed 12 trials of telemonitoring in patients with CHF. Two of the studies analyzed by Giamouzis et al. were included in the review by Clark et al.. The review reported on these studies, discussing their characterestics and the significance of their findings. We note that although some of the reviewed studies by Giamouzis et al. collected self-reported physical activity data, no reviewed study collected quantitative activity data (such as energy expenditures (EE)) [Giamouzis et al., 2012].

In a more recent randomized clinical trial, Ong et al. [2016] reported no significant differences in 180 or 30-day readmission rates or 180-days mortality rates. The trial was conducted with 1437 patients with CHF, where 715 patients were assigned to the treatment group and 722 were assigned to the usual care group. The patients in the treatment group were subject to interventions that combined health coaching telephone calls and telemonitoring through electronic equipment that collected daily information about blood pressure, heart rate, symptoms and weight. The study further reported significant improvements in 180-days quality of life for the patients in the

treatment groups [Ong et al., 2016].

We note from our literature review, that although physical activity is an important marker in assessing the risk of clinical deterioration in CHF, no study from the surveyed above included a continuous and quantitative way to remotely monitor patients' physical activity levels. We review various means of continuously measuring of physical activity, that can be applied to mHealth telemonitoring systems in Section 2.2.3. We also note that many of the studies surveyed above relied on structured telephone support to accomplish telemonitoring, perhaps indicating the difficulty to build reliable means of autonomous telemonitoring, which is *the* issue we address in this chapter.

2.2.2 Mobile Health (mHealth)

In this chapter, we are interested in the design of telemonitoring systems in general, and those that are implemented on top of mobile systems in particular. Therefore, we briefly review the field of mHealth and some of the advancements in smartphone-related technologies related to health and fitness monitoring and coaching. Similar to the field of clinical telemonitoring in patients with CHF, this field is also very comprehensive in contributions. Therefore, we elected to summarize the findings of a comprehensive review by Hussain et al. [2015], which included 133 articles in the field of mHealth from IEEE Xplore, MEDLINE, ScienceDirect and Web of Science [Hussain et al., 2015]. Interested readers are referred to the paper and the references therein for more details.

Hussain et al. surveyed the trends of using smartphone apps for clinical, medical and fitness purposes, since the beginning of 2010. They found that the majority of articles in the field (68 out of 133) either addressed specific medical apps or provided an overview of apps dedicated to a specific disease area, or a specific clinical specialty or tool. Examples of these specific areas include i) anesthesia, ii) asthma, iii) cardiology, iv) cardiopulmonary resuscitation (CPR), v) dentistry, vi) dermatology, vii) endocrinology, viii) family medicine, ix) infectious diseases, x) internal medicine, xi) oncology, xii) ophthalmology, xiii) palliative medicine, xiv) pediatrics, xv) pharmacy, xvi) psychiatry, xvii) public health, xviii) rehabilitation, xix) sports medicine. xx) surgery (including plastic surgery) and xxi) women's health. It is clear, from the list above, that the field of mHealth is gaining traction and is becoming ubiquitous in terms of its applications.

Hussain et al. identified that a large group of surveyed articles (43 out of 133) were concerned in reporting on the design aspects and usability of clinical, medical and fitness apps. The methods used in these articles included evaluating existing apps or identifying requirements and features of mHealth applications. In a similar category, 17 articles (out of 133) were found to report on the lessons learned in the process of developing and implementing new clinical, medical and fitness apps. The rest of the reviewed articles (5 out of 133) were focused on general frameworks aiming to address mHealth development and operation.

To situate this chapter in the aforementioned categories, it essentially fall under all of these categories. We discuss specific medical and fitness apps for a specific purposes–the CHF study app and the RunningCoach app. We report on the design and usability of these apps and we discuss the lessons learned in the process of developing these apps. Finally, we focus on systematically addressing the challenges faced during the design of these apps in a general mHealth framework, the Berkeley Telemonitoring framework.

Finally, we note that Hussain et al. emphasized in their report that i) the development of medical apps is not standardized, resulting in a relatively small number of researchers developing medical apps; and ii) medical apps need clinical validation. We also believe that mHealth systems need to be standardized in order to ignite further progress in the field. Furthermore, we believe that standardization will assist in the effort of clinical validation. In Section 2.5, we discuss several

commercial platforms that attempt to standardize the development and operation of mHealth systems. Afterwards, we outline the outstanding challenges and introduce our contribution to the efforts of standardization, the Berkeley Telemonitoring project.

2.2.3 Quantitative Measurement of Physical Activity

Physical activity is an important factor in preventing several chronic diseases, including CHF, hypertension, diabetes, obesity, cancer, osteoporosis and depression [Warburton et al., 2006]. From our review of studies of telemonitoring in Section 2.2.1, it was evident that many studies don't collect objective data on physical activity. The few studies that attempted to collect some form of physical activity data, did so through self-reported means. This approach, as simple as it is, suffers from major issues when it comes to the reliability of the collected data. For instance, an expert report by Dhurandhar et al. [2014] argued that such self-reported data on physical activity are too inaccurate to be used for clinical purposes [Dhurandhar et al., 2014].

As such, it is vital for telemonitoring systems to incorporate means for the collection of objective measures of physical activity. It has been demonstrated in a range of health and fitness applications that an objective measure of physical activity has the potential to encourage individuals to initiate and maintain a healthy lifestyle for a longer periods of time [Klaassen et al., 2013]. In general, physical activity can be quantified by estimating EE. In this section, we review the different efforts in the field of measuring activity levels quantitatively. One line of research efforts attempts to incorporate physical activity sensing capabilities in daily-use products such as clothing [Axisa et al., 2003; Lee and Chung, 2009].

In the commercial domain, various products for EE data collection are available for use. Some examples include Fitbit Activity Wristbands and Trackers[1], Nike+ FuelBand[2] and others. In the realm of validating these devices, multiple research studies argued that many of these devices are accurate only in step-counting, but are inaccurate in estimating EE [Pande et al., 2013; Dannecker et al., 2011]. In a more recent study, Case et al. found that smartphone apps can be more accurate even in counting steps than some of the wearable devices [Case et al., 2015]. We conclude that it is vital to include sensor-accuracy models in any system for predictive medicine such as telemonitoring.

In addition, many of the commercial wearable devices for EE measurement require the use of a smartphone. Therefore, if they are to be used in telemonitoring systems, their use would likely not eliminate the use of a smartphone. Since smartphones are now ubiquitous (at least in the developed world), we argue that it is feasible to achieve an objective and accurate continuous measurement of EE using a smartphone only [Patel et al., 2015]. With that in mind, it is still important to design such a measuring module without incurring a heavy cost on battery life, because it would otherwise be a limiting factor in adopting such a technology [Alshurafa et al., 2015].

In particular, most (if not all) modern smartphones are equipped with accelerometers that are used in a range of applications and studies for the purposes of physical activity monitoring, recognition and classification [see Donaire-Gonzalez et al., 2013; Pande et al., 2013, for example]. These applications and studies have provided evidence to the reliability of EE estimation using a smartphone only [Pande et al., 2013]. In the next section of this chapter, we describe a study that we conducted, in collaboration with NMFF, in order to understand the requirements of telemonitoring systems [Aranki et al., 2016b]. In this study, we included continuous EE estimation

[1] `http://www.fitbit.com/`

[2] `http://www.nike.com/us/en_us/e/nike-plus-membership`

using a smartphone only; an algorithm developed by Chen and Sun [1997] and validated by Donaire-Gonzalez et al. [2013]. The aforementioned algorithm relies entirely on the built-in accelerometer sensor available on the smartphone.

2.3 Intervention

One of the key aspects of the predictive healthcare model is the ability to act based on the predictions made. Therefore, we now turn to discuss the notion of medical intervention. There are two main categories of medical intervention. First, health-behavioral interventions are intended to motivate patients to start and maintain health-improving activities. The second category of interventions are preventive interventions, which are geared towards discouraging behaviors that are detrimental to the patient's health [Spring et al., 2013]. The origins of health-behavioral interventions stem from theories in social sciences (c.f. positive reinforcement). These theories relied, for the most part, on observational studies and self-reported data.

The type of medical intervention relies on the health-risk assessment that is extracted from the data by using health-behavioral and predictive models [Chih et al., 2014]. Because of that, monitoring is key to the efficacy of the intervention. this monitoring should not only consist of initial observations to initiate medical intervention, but also to monitor the effects of the intervention for further refinement. Spring et al. [2013] characterized this process by *4 Ms*: monitoring, modeling, motivating, and modifying [Spring et al., 2013].

We argue that real-time telemonitoring through mobile technologies enables us to revisit the notion of health-behavioral modeling. Now more than ever, we are able to collect objective estimates of health-related parameters, such as vital signs, EE and phone usage. In addition, we can fuse these objective measurements with more subjective pieces of information, such as emotional/mental state, pain and energy level. This mixture can help us better understand and construct health-behavioral and predictive models, and better assess the effectiveness of a given medical intervention. It is important to note that such models have to be dynamically adapted since the patient's behavior changes over time. Given this dynamic nature of behavioral changes, motivation becomes a stronger aspect of changing and maintaining healthier behaviors [Spring et al., 2013]. Including aspects from personalized medicine are also important in the intervention process This is because each individual is motivated in a particular way.

Evaluating health and medical interventions as delivered through smartphones has been the subject of various studies in various applications. These applications include i) physical activity [Burns et al., 2011]; ii) depression [Burns et al., 2011]; iii) promoting weight loss [Martin et al., 2015]; and iv) schizophrenia [Ben-Zeev et al., 2015].

2.4 CHF Study

2.4.1 Introduction

Given the vast potential of telemonitoring, we wanted to better understand the requirements and usability of such a technology, and elected CHF as a field with a large potential benefit. Therefore, in 2012, we designed a pilot study for telemonitoring in patients with CHF, for that purpose. Moreover, given the lack of telemonitoring studies collecting objective measures of physical activity, we decided to include a collection of EE estimates in the telemonitoring application. This study,

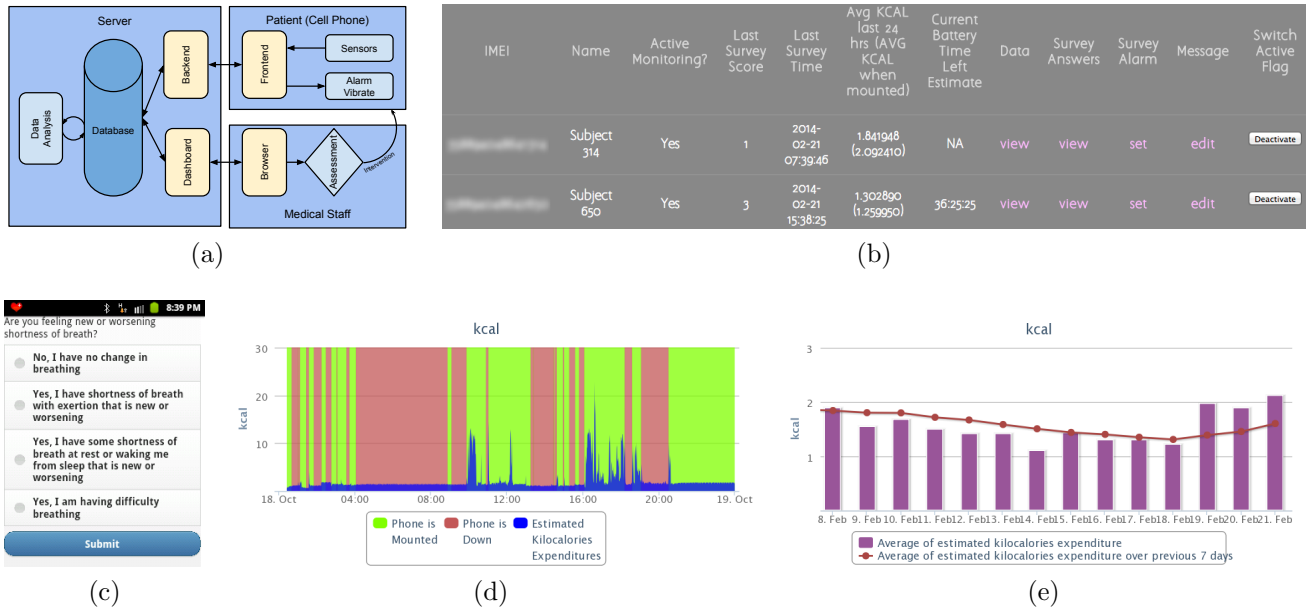


Figure 2.2: (a) The architecture of the CHF study system; (b) the CHF study dashboard main menu; (c) The CHF telemonitoring app; (d) minute-by-minute EE estimates, overlaid by the classified state of the phone; and (e) daily average EE estimates, bars depict daily averages while the line depicts moving 7-day averages.

to our knowledge, was the first in the field of telemonitoring in patients with CHF to collect continuous measurements of EE.

The architecture of the designed telemonitoring system is depicted in Figure 2.2a. Note that the system design conforms to the workflow of the predictive healthcare model, as depicted in Figure 2.1. In short, the workflow of the designed system consists of i) data collection, using a smartphone telemonitoring app (EE, vital signs and symptoms); ii) submission of the collected data from the smartphone to a central server; iii) analysis of the data to provide risk assessment, using a simple predictive model designed by the cardiologists in the team; and iv) medical intervention, delivered through the telemonitoring app.

The telemonitoring app was designed for Android phones running API level 9 or above (Android 2.3 GINGERBREAD or above).[3] A screenshot of the app is depicted in Figure 2.2c. The backend consisted of a server running a MySQL database for data storage. A screenshot of the backend's dashboard, available to the medical staff, is depicted in Figure 2.2b. The communication between the telemonitoring app and the server was carried over Secure Sockets Layer (SSL) for encryption. In this section, we summarized the challenges faced in the design of this system and the findings of the study. We will elaborate more on the privacy and acceptability aspects of this study in Section 3.2.2. For more details on this study, we refer the reader to [Aranki et al., 2016b].

The study was approved by the Institutional Review Board at NMFF and was carried out between July 2013 and May 2014. During that time, 34 patients with CHF were approached for enrollment in the study, and 15 patients accepted to participate. The demographics of the participating subjects are reported in Table 2.1. Our protocol was designed to monitor each patient for a period of 3 months. In practice, some patients elected to participate for a longer period of time, and others elected to quit the study before the end of that period.

[3] `https://developer.android.com/guide/topics/manifest/uses-sdk-element.html`

		Count
Gender	**Male**	8
	Female	7
Race/Ethnicity	**Black or African-American**	12
	White, non-Hispanic	3
Education	**High-school or less**	3
	Beyond high-school, <4 years of college	8
	4 year college graduate	1
	Graduate school	3
Household income	**Prefer not to say**	5
	$<\$5000$	2
	$\$5000 - \$14999$	4
	$\$20000 - \$49999$	2
	$\$50000 - \$59999$	1
	$\$60000 - \$99999$	1

Table 2.1: The demographics of the subjects.

2.4.2 Collected Data

To give context to the reporting on this study, we first iterate over the different types of data collected in this study. The first category of collected data is health-related data. In this category, we collected minute-by-minute EE estimates, based on an algorithm developed by Chen and Sun [1997]. Example plots of EE data are depicted in Figures 2.2d and 2.2e. Moreover, we collected, in a self-reported fashion, daily information about the subjects' symptoms (dizziness, fatigue, shortness of breath, chest discomfort and activity level) and vital signs (heart-beat rate, blood pressure and weight).

In order to better understand the usability of the telemonitoring app, we also collected minute-by-minute phone-related data such as the battery level, whether the phone is charging and whether the screen light is on. The list of variables collected in this study are listed in table Table 2.2.

Among other things, the phone-related data were used to assess the status of the phone in order to tag the EE data for reliability. That is, we were interested to be able to tell when the phone is actually on the subject's body versus when it is not, in order to know when the EE data are reliable. The health-related data, on the other hand, were used to assess the level of risk of clinical deterioration.

This assessment was done through calculating a risk score based on the health-related data, once a day. The design of this risk score formula was done by cardiologists who were part of the research project. The risk score captures the presence and severity of the symptoms collected [Remme and Swedberg, 2001; Swedberg et al., 2005]. It also captures the absolute values of systolic blood pressure and heart-beat rate, and the relative weight compared to the day before.

Depending on the range of the resultant risk score, the patients received a message, daily. The messages ranged from encouraging the patient when the estimated risk is low to urging the patient to take immediate action, including calling 911, if the estimated risk is too high. The way the risk score calculation was designed, the higher the score the higher the estimated risk of clinical deterioration is. The different messages of intervention, depending on the different ranges of risk scores, are reported in Table 2.3.

Type	Category	Description
Sensory, minute-by-minute	Health	EE estimate (aggregated)
	Phone	Tri-axial magnetic field (averaged)
		Gravity pointer (averaged)
		Tri-axial phone orientation (averaged)
		Tri-axial phone rotation (averaged)
		Battery level (averaged)
		Screen light (% of minute with light on)
		Proximity (% of minute with an object close to the phone)
		Call status (% of minute a call was in session)
		Charging status (% of minute with phone charging)
		GPS, longitude is shifted
Self reported, once a day	Symptoms	Level of fatigue
		Level of shortness of breath
		Level of dizziness
		Level of chest discomfort
	Activity	Level of activity
	Vital signs	Heart-beat rate
		Blood pressure
		Weight

Table 2.2: Data collected by the CHF telemonitoring app.

2.4.3 Study Findings

In this section, we briefly report on the findings of the CHF study. The full details of these findings can be found in [Aranki et al., 2016b]. After enumerating the study findings, we will summarize them, in conjunction with the challenges faced while designing the telemonitoring system. This summary (Section 2.4.4), in essence, serves as a set of lessons learned that will guide our design of the general telemonitoring framework, the Berkeley Telemonitoring framework.

Reaction to Intervention One of the findings of the study was that subjects' behavior as a result of intervention is neither uniform nor static. In other words, different patients react to the same intervention differently under similar circumstances, and the same patient may react differently to the same intervention. In the case of this study, we observed that some subjects complied and called for medical attention when prompted to do so, while others did not. Moreover, we observed, in the case of one subject, that she or he did not originally take the suggested action in the intervention. The patient was readmitted to the emergency room shortly after her or his risk score came too high. Afterwards, the same subject started better following the suggested action in the intervention after high risk scores. These observations are in line with our discussion from Section 2.3.

Adoption and Usage Fatigue Another interesting result from the CHF study is that the patients' willingness to use the technology wears off with time. We refer to this phenomenon by *usage fatigue*. To be precise, we consider usage fatigue to be the decrease in the proper usage of the technology over time. In the study, we devised a quantitative measure for usage fatigue and used it to quantify this phenomenon.

Score Range	Feedback Text
≥ 4	Please page the on-call cardiologist at 312-695-XXXX or call 911 for immediate attention if needed. Your responses suggest a need for quick evaluation.
3	Please contact the Heart Failure Clinic at 312-695-XXXX. Your responses suggest the need to check in now.
$1-2$	Thank you for your response! If you are feeling worse, please contact the Heart Failure Clinic at 312-695-XXXX.
≤ 0	Thank you for your response! Your results appear stable, overall. Please remember your diet and exercise goals!

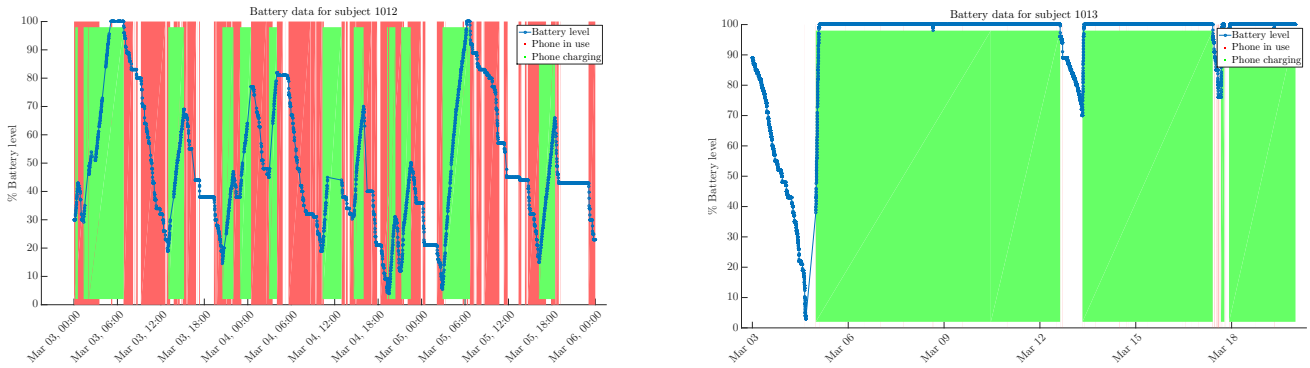
Table 2.3: The intervention message, depending on the estimated risk of clinical deterioration.

Data Plan Consumption Identifying that excessive cellular data plan consumption may be a prohibitive factor in adopting an mHealth technology, we designed the telemonitoring to be mindful of its use of cellular data. We validated our design in a laboratory test, which had the following setting. The telemonitoring app embedded packets of 30 data points per transmission. Each data point is 1 minute worth of telemonitoring data. The size of each transmission (including failure recoveries) had a mean of 19.15 kB and a standard deviation of 1.6 kB. There are at most 1488 packets a month (assuming full-time monitoring for 31 days). If we assume all transmission sizes are independent, we get that the monthly data consumption (sum of all individual transmission sizes) is normally distributed with a mean of 27.82 MB and a standard deviation of 2.33 MB. From this we conclude that the probability of consuming more than 35 MB in a month is less than 1%.

Battery Consumption Battery consumption is a key aspect for adoption. During the study, some subjects complained about short battery life due to the telemonitoring app. Other subjects reported no battery consumption issues when asked. It is important to mention i) that all subjects were provided the same phone make and model by the study; and ii) that we had tested battery consumption before the deployment of the study. During these tests, we found that the phone's battery consistently lasted more than 20 in standby mode. Our hypotheses for the reported high battery consumption issues included i) low cellular signal coverage, which increases battery consumption; ii) multiple sessions of non-full battery recharging a day may give the impression of high battery consumption; and iii) high usage of the phone for purposes that are not related to the study (such as, surfing the web or playing multimedia). As examples of the different charging and usage behaviors, we display two of the patterns of charging and battery consumption from the subjects of the study in Figures 2.3a and 2.3b. "Phone in use" indicates that the phone is currently being actively used (either a phone call, or the screen is on).

2.4.4 Challenges and Lessons Learned

Proper Authorship Perhaps the largest leap in designing ubiquitous systems in general, and telemonitoring systems in particular, is that they operate in real-world conditions. That is, the users will use these technologies in their home, commute, workplace, etc. Many of the assumptions made during the design phases of these systems are primarily validated in laboratory conditions only. For example, in this study, the EE estimation algorithm used in our system assumes that



(a) (b)

Figure 2.3: Excerpts of battery recharging and consumption patterns of (a) subject 1012 over the course of three days; and (b) subject 1013 over the course of 17 days.

the phone is worn at the waistline, on the right hip [Chen and Sun, 1997; Donaire-Gonzalez et al., 2013].

When the system is deployed, however, this assumption needs to be validated, so that the reliability of the collected data can be assessed. For example, if the phone is left on the table, the resulting EE estimates should not be trusted. This is particularly important when such data are used to assess risk of hospitalization or take decisions about medical intervention. Therefore, we argue that *proper authorship* is a challenge that needs to be addressed systematically. For instance, when designing a vital sign estimation algorithm, we need to design validation algorithms that can detect when the design assumptions of the estimation algorithm are violated.

Compliance and Usage Fatigue As discussed earlier, patients' *compliance* is another challenge in deploying telemonitoring systems. Their willingness to use the technology tends to decrease over time. We refer to this phenomenon by *usage fatigue*. We observed that patients started using the technology less and less as time passed by. From the subjects' feedback in this study, it became clear that the technology is passive. Therefore, we argue that a better understanding of *incentives*, *user interfaces* and *user feedback* can greatly benefit long-term compliance for telemonitoring systems.

Reliability and Objectiveness of Data Health-related markers that are to be used in risk assessment and intervention decision making have to be as *reliable* and *objective* as possible. This resonates with the expert report published by Dhurandhar et al. [2014], which argued that subjective, self-reported data on activity levels are too unreliable for clinical use. Instead, telemonitoring systems have to incorporate estimation algorithms that rely on sensory data. Moreover, *reliability and objectiveness models* are vital, particularly if these systems are to be scaled for a large number of patients. These models have to be incorporated in the predictive models that are responsible for the intervention decision making process.

Physio-behavioral Models During the study, we observed that the reaction of patients to medical intervention is neither uniform nor static. As discussed earlier, this means that different subjects may react differently to the same intervention even under similar circumstances. Moreover, the same subject may behave differently, in response to a medical intervention, under different circumstances.

Therefore, models that rely solely on physical parameters are not sufficient for predictive models and the intervention decision-making process. In addition to physical parameters, behavioral markers should be incorporated into these models and processes. We refer to these models as *physio-behavioral models*. These models can greatly assist in increasing the efficacy of medical intervention, as discussed in Section 2.3. These models are not trivial to construct and validate and will require large behavioral studies.

In summary, in order to scale telemonitoring systems, manual intervention decision making will not be sufficient, and autonomous expert systems have to replace them. For those to work effectively, physio-behavioral models have to be developed.

Privacy Perhaps one of the most challenging aspects of developing health technologies, and certainly one of the ethical issues involved, is the protection of *privacy*. Telemonitoring technologies have to be minimally intrusive to the monitored user *and* to other individuals this user interacts with (e.g., co-workers and family members). That is, telemonitoring should in general only collect, analyze and retain data in order to achieve its set medical goals, and not more.

We believe that a firm understanding of the privacy requirements of telemonitoring systems and implications of their design is vital to its ethics and success. We discuss this challenge in great detail in Chapter 3.

2.5 Commercial Solution

2.5.1 Introduction

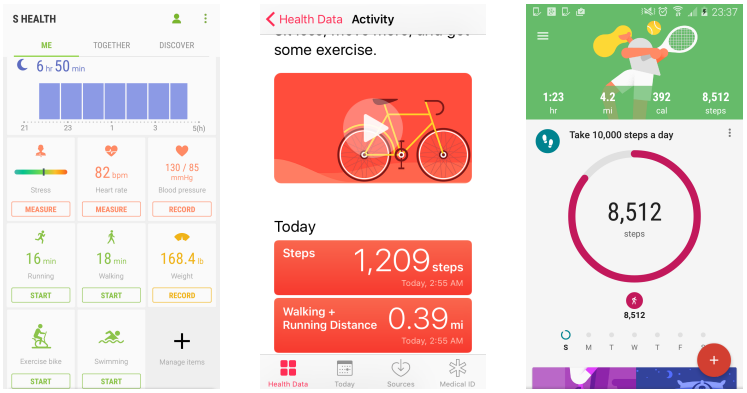
As a result of the CHF study, we now turn to systematically address the challenges faced in its design and the lessons learned as a result of conducting it. Before we move to describing our solution to standardizing mHealth telemonitoring, we first survey the field for similar attempts to standardize mHealth and/or telemonitoring. In Sections 2.5.2 to 2.5.5, we briefly describe the key platforms in this realm and then move to summarizing the challenges that remain open in Section 2.6. This will provide us the necessary context to describe the Berkeley Telemonitoring project and its resultant framework.

2.5.2 Samsung Digital Health and S Health

In July 2012, Samsung released their health platform *S Health* alongside the release of their Android phone, Galaxy S III. S Health was later renamed to Samsung Health in April 2017. Originally, their intention was to provide a compatible wellness platform that is able to communicate with blood pressure monitors, glucose meters and body composition scales. Samsung later released the *Samsung Digital Health* software development kit (SDK) to third-party developers for the development of wellness apps, compatible with Samsung Health.

The core features of Samsung Health are as follows. First, the platform provides mechanisms for sharing data between sensors and various consumer applications for a multitude of uses, including coaching, and social interaction. In its dashboard, depicted in Figure 2.4a, Samsung Health summarizes the data obtained from the various wellness apps and provides different visualizations including graphs and tables.

Samsung Health was originally limited to Samsung Galaxy devices. However, Samsung extended the support of Samsung Health in September 2015 to include all devices running Andorid



(a) (b) (c)

Figure 2.4: (a) A screenshot of the Samsung Health summary screen; (b) a screenshot of the Apple Health activity tracking screen; and (c) a screenshot of the Google Fit activity tracking screen.

4.4 (KitKat) or newer. Moreover, Samsung has built multiple collaborations with academic researchers in the field that aim to "accelerate the validation and commercialization of new sensors, algorithms, and digital health technologies" for preventive medicine.[4] For example, in 2014, Samsung and University of California San Francisco (UCSF) establishing the UCSF-Samsung Digital Health Innovation Lab. As part of the announcement of establishing this lab, Samsung released their open hardware platform *Simband*, which aims to enable developers to design their own wearables that include measurements of electrocardiogram (ECG), heart-beat rate and other variables. These sensors were also being validated in collaboration with UCSF [Bloss, 2015].

As for the server side, Samsung Health takes advantage of Samsung ARTIK Cloud [5] (formerly known as SAMI), which would enable large-scale clinical trials and epidemiological studies. This framework can provide researchers with capabilities to collect, securely store, view, and analyze data in real-time. However, at the time of writing this dissertation, Samsung does not have a platform designed for clinical trials specifically.

2.5.3 Apple Health, ResearchKit, and CareKit

In September 2014, Apple released the *Health* app, with the *HealthKit* application programming interface (API). This is, in a way, Apple's response to Samsung Health. Apple Health provides users access controls for allowing or denying apps from accessing health and fitness data. Apple Health is designed to collect health and fitness data including calories burned, blood pressure, heart-beat rate and blood sugar. Similar to Samsung Health, Apple Health provides a dashboard, depicted in Figure 2.4b, that summarizes the collected data and visualizes them in a user-friendly manner.

As opposed to Samsung, Apple did release an open-source framework, *ResearchKit*, for the collection of health-related data targeting clinical and medical research [Apple Inc, 2016]. ResearchKit was released in March 2015, with the release of Apple Watch. The framework is designed to enable researchers to develop their own apps for data collection from i) surveys and self-reported vital

[4] `https://www.ucsf.edu/news/2014/02/111976/samsung-ucsf-partner-accelerate-new-innovations-preventive-health-technology`

[5] `https://artik.cloud/`

signs; ii) Apple Health, provided the user's permission; and iii) movement, based on the phone's internal sensors and/or wearable devices [Jardine et al., 2015]. Given the ubiquity of smartphones in the developed world, ResearchKit aims to streamline health-related data collection in an effort to accelerate medical research, such as epidemiology.

With the release of ResearchKit, Apple simultaneously announced five studies that utilize ReseachKit-built apps. The first apps based on ResearchKit were (1) *mPower* app (University of Rochester and Sage Bionetworks) for tracking symptoms in Parkinson's Diseases, (2) *GlucoSuccess* app (Massachusetts General Hospital) to track diet, physical activity, and medication in persons with diabetes, and (3) *Asthma Health* app (Mt. Sinai, Weill Cornell Medical College, and LifeMap) for collecting asthma related markers, (4) *Share the Journey* app (Dana-Farber Cancer Institute, UCLA Fielding School of Public Health, Penn Medicine, and Sage Bionetworks) to study quality of life of patients treated for breast cancer, (5) *MyHeart Counts* app (Stanford University and American Heart Association) to study risk factors for cardiovascular disease [Taylor, 2015]. More studies have been announced since then, including *CTracker* that aims to understand the impacts of hepatitis C in daily life, the *Autism & Beyond ResearchKit* app for autism and the *Mind Share* app that aims to study Alzheimer's disease.

An important aspect of ResearchKit is its ability to provide an electronic version of consent for research studies. We will elaborate further on the consent issues in general in Chapter 3, but in essence, this enables ResearchKit to allow researchers a greater outreach to subjects. For example, within 24 of its release in March 2015, the MyHeart Counts app gained over 10,000 participants [AppleInsider Staff, 2015]. On the other hand, this has raised a multitude of ethical concerns, including research on minors, privacy and potential issues with the consent being an informed consent process [Hunter, 2015]. It is important to note that even though ResearchKit provides encryption for data storage and transmission, the onus of complying with health regulations–such as Health Insurance Portability and Accountability Act (HIPAA) in the United States of America (US)–falls on the researchers conducting the study [Ritter, 2015].

Finally, Apple released *CareKit*, in March 2016, an open-source framework targeting self-management of diseases by the patients. CareKit enables developers to build apps that allows patients to actively monitor their conditions and share relevant data with their caregivers. Two of the early examples of apps designed using CareKit include i) *Parkinson's Central* (by National Parkinson Foundation, Inc.) that allows patients with Parkinson's disease and their physicians to track symptoms and medications; and ii) *One Drop* (Informed Data Systems, Inc.) app for diabetes management. More recently, other apps that are built over CareKit have been announced for conditions including mental health management, post-surgery progress management and maternal health management.

2.5.4 Google Fit

In October 2014, Google released *Google Fit*, a health-tracking platform for the Android operating system Similar to Samsung Health and Apple Health, Google Fit addresses the collection and aggregation of health and fitness-related data from popular trackers and health-related apps. Google Fit utilizes both internal smartphone sensors and external wearable devices to count steps and classify exercise activities, such as running, walking and cycling. Google Fit also provides a dashboard, depicted in Figure 2.4c, which allows the user to view her or his data in a centralized location by visualizing the data in a user-friendly way.

Ensuing Samsung and Apple's announcements regarding their interest in clinical research, Google announced, in June 2015, that its research division has developed a health-tracking wrist-

band that is designed for clinical and drug trials [Chen and Womack, 2015]. The wristband is designed to measure minute-by-minute markers including skin temperature, heart-beat rate, noise levels and light exposure. At the time of writing this dissertation, no further information has been released regarding this effort.

2.5.5 Other Commercial Platforms

Several other (and smaller, in market share) commercial mHealth frameworks exist in the market. In October 2014, Microsoft released *Microsoft Band* and *Microsoft Health*, the company's wearable device and health-tracking platform, respectively. Microsoft Band, in conjunction with Microsoft Health, allows users to collect and view their fitness and health data, including sleep quality, workouts and step counts. Since 2014, Microsoft Health has broadened its support to include other wearables. Moreover, Microsoft Health is now able to connect to partner apps, such as i) *MyFitnessPal*, ii) *Strava*, iii) *RunKeeper*, and iv) *MapMyFitness*. Microsoft released an SDK for developers, that takes advantage of the Microsoft Health Cloud API. This, in turn, allows developers to access the measurements collected by the smartband by utilizing Microsoft's cloud and the tools therein.

Another examples of a mHealth and mobile fitness platform is Under Armour's *Under Armour Connected Fitness*[6] and its app *UA Record*. This platform enables the collection of data related to activity, workouts, and sleep from third party devices and internal smartphone sensors. Under Armour Connected Fitness also includes social networking features that enable users to share and compare their data with their friends, as well as participating in customized fitness and health challenges. This platform is accompanied by an SDK for developers. In addition to the software platform, Under Armour also released a hardware bundle, called *UA HealthBox*, which includes a wristband, a scale and a chest heart-beat rate monitor that allows users to collect sleep, nutrition and activity related data.

Boiling down to apps with specific purposes, several wearable manufacturers designed apps that aim to extract data from their wearables, enabling users to track their progress over time, and share collected data with other users. Although these apps are standalone apps, some of them can also connect to all or some of the major framework, such as Samsung Health, Apple Health and Google Fit. Three such examples are i) FitBit[7], ii) Nike FuelBand[8], iii) Pebble watch[9], and iv) JawBone[10].

A taxonomy summary of the different mHealth frameworks presented in this section can be found in Table 2.4. We include the Berkeley Telemonitoring framework that will be presented in Section 2.7, for completeness. Next, we survey the open challenges in the field of mHealth in general, and mHealth telemonitoring in particular, providing context to our framework, the Berkeley Telemonitoring framework.

2.6 Challenges

The commercial platforms and frameworks described in Section 2.5 constitute a step forward towards achieving the full potential of mHealth telemonitoring. However, many challenges remain

[6] `https://www.underarmour.com/en-us/ua-record`
[7] `https://www.fitbit.com/`
[8] `https://www.nike.com/us/en_us/c/nike-plus`
[9] `https://www.pebble.com/`
[10] `https://jawbone.com/`

Framework	Target use	Features and contributions	Disadvantages
Samsung Health	Consumer health & fitness tracking.	Android-based platform for data collection and aggregation from health apps and various sensors. Includes tools for data analysis, social interaction and coaching with security security and encryption integration.	Limited to Android OS.
Apple HealthKit	Consumer health & fitness tracking.	Apple iOS based platform for health data aggregation, visualization and exchange from apps and external sensors. Features information sharing with health providers and promises future integration with electronic medical records.	Lacks support for other mobile platforms (Andorid, Windows).
Apple ResearchKit	Medical research; Clinical trials.	Open-source framework for large-scale medical research data collection. Provides modules for i) informed consent; ii) surveys; and iii) active tasks that include motor activities, audio, cognition, etc.	Limited to Apple iOS, there is potential population bias [Jardine et al., 2015]. Potential informed consent issues [Hussain et al., 2015; Hunter, 2015].
Apple CareKit	Patient self-management.	Open-source framework for patient-centric disease self-management apps development. Features information sharing with caregivers, doctors and family.	Lacks support for other mobile platforms (Andorid, Windows).
Google Fit	Consumer health & fitness tracking.	Android-based platform for data collection from popular health-related apps and fitness trackers. Features high-level representation of sensor data, fitness data types and training sessions.	Primary focus is on fitness data, less on health data. Privacy concerns regarding data sharing with 3rd parties.
Microsoft Health	Consumer health & fitness tracking.	Cross-platform (Windows, Android, iOS) health-tracking framework with support for popular health-related apps and trackers. Features data aggregation, storage and analytics with cloud integration.	Not widely popular.
Under Armor Connected Fitness	Consumer health & fitness tracking.	Platform for aggregating data on activity, workouts and sleep from devices and sensors (including third-parties). Features social networking components for fitness users and athletes.	Primary focus is on fitness data, less on health data.
Berkeley Telemonitoring	Medical research; Clinical trials; Consumer health & fitness tracking.	Android-based open-source framework for development of health-monitoring apps. Provides libraries for clients and servers. Features distributed computation, secure and robust data storage, and modules for survey delivery. Features connectivity to medical devices and several validated algorithms for estimation of vital signs.	Limited to Android operating system.

Table 2.4: Taxonomy summary of the different health and fitness frameworks [Aranki et al., 2017b].

open in this realm. In this section, we describe some of these challenges, in order to give context to the Berkeley Telemonitoring project and its framework (Section 2.7).

Clinical Challenges There is currently a lack of involvement of qualified healthcare professionals in the design, development and deployment of mHealth systems in general, and telemonitoring systems in particular [Hussain et al., 2015]. As a result, there is also a lack of validation for the clinical efficacy and effectiveness of such systems. Moreover, more research in evaluating the clinical outcomes of mHealth systems is also needed [Eng and Lee, 2013; Aranki et al., 2016b]. Most research projects focus on the benefits and positive effects of mHealth systems, but not as many focus on the negative effects of such systems. These studies can greatly contribute to building predictive models that incorporate behavioral trends and markers (what we called physio-behavioral models in Section 2.4.3), which are necessary for the effectiveness of mHealth systems [Eng and Lee, 2013; Hussain et al., 2015; Aranki et al., 2016b].

Even though there seems to be a vast amount of mHealth systems, frameworks and devices that enable the collection of data, there is a noticeable gap in clinically verified autonomous expert systems that can provide timely feedback and decide on clinical intervention, based on the such data [Aranki et al., 2016b,a]. Particular to mobile-based systems, but also applicable to general ubiquitous-like telemonitoring systems, there is a challenge of validating that the data indeed apply to the intended monitoring subject, and not someone else who happens to be in the vicinity of the telemonitoring system (we referred to this challenge by the challenge of proper authorship). Tools that can enable us to identify such scenarios and avoid them are needed [Aranki et al., 2016b]. We provide more context to this challenge in Chapter 3.

Systems and Standardization Challenges There is a need for a regulatory framework that standardizes the design and development of telemonitoring and mHealth systems [Hussain et al., 2015]. The lack of standardization results in a noticeable change in interfaces between mHealth platforms and existing healthcare systems, such as electronic health records systems and databases [Hussain et al., 2015]. The lack of smooth integration, particularly in user interfaces, results in a bad user experience that affects the usability of mHealth systems for patients. These challenges include (a) patients interest in using these systems wears off over time (we called this challenge "usage fatigue" in Section 2.4.3); (b) patients have to climb through a learning curve with each new mHealth systems; and (c) lack of seamless integration creates a bias against patients in rural areas due to poor cellular coverage [Hussain et al., 2015; Aranki et al., 2016b]. For example, in practice, a roadblock in adopting such systems by healthcare providers is the lack of integration with existing reimbursement and healthcare delivery systems; a problem that can be, in part, due to the lack of standardization [Eng and Lee, 2013; Hussain et al., 2015].

As a result of many of these challenges, most mHealth frameworks and systems, including telemonitoring ones, enable consumers to collect data related to a specific task, but do not deliver standard medical quantities that are based on the collected data. As an example, many frameworks allow developers to estimate step counts from accelerometer data, but do not provide estimation algorithms that output medical quantities such as EE based on the same data. Most often the time, whenever a commercial platform supplies such an estimation ability, they do it in a proprietary manner, which results in i) making its use limited to the company's products; and ii) making it hard to clinically validate and compare different means to estimate the same medical parameter due to lack of transparency. This lack of transparency and standardization also limits the unification of data collection and sharing data between different studies that use different estimation mechanisms

because it would otherwise hinder drawing reliable conclusions if one is not careful. These problems are not confined only to estimates of complex health parameters and are present in even simpler ones like step counting [Case et al., 2015].

Legal and Ethical Challenges Some of the challenges in deploying an effective mHealth framework span beyond scientific or technological boundaries. For instance, the delivery of informed consent through mobile means is a challenging aspect of designing mHealth technologies, specially for clinical research purposes. This challenge is even magnified when these clinical studies target a large number of subjects, as is the case with those conducted using Apple ResearchKit [Hussain et al., 2015; Hunter, 2015]. As of July 2017, a clinical study by Duke University is being designed aiming to compare the standard informal consent process to the modular consent process offered in Apple ResearchKit.[11] There are also concerns, in the scientific community, that some mHealth frameworks, such as Apple ResearchKit, come with an inherent bias due to the demographics of people using specific smartphones in general, and specific smartphone platforms in particular (such as Apple iOS, in the given example) [Jardine et al., 2015]. We elaborate more on some of these issues in more detail in Chapter 3.

Administrative Challenges When developing new mHealth systems or interfacing them to existing health systems, the financial cost is often a discouraging factor that impedes their development and adoption [Hussain et al., 2015; Boulos et al., 2014]. This magnifies the need for standardization in the field, which may allow faster and cheaper development and integration of such systems, thus alleviating this challenge. Security and privacy-protection constitute another challenge in deploying mHealth systems. Hussain et al. argue that there is little research being done on privacy of mHealth systems relative to the amount of research being carried on the privacy of more traditional healthcare technologies, such as electronic health records [Boulos et al., 2014; Hunter, 2015; Hussain et al., 2015; Aranki et al., 2016b]. Most mechanisms currently adopted in mHealth systems to protect consumer privacy revolve around access control and granting or revoking permissions (e.g., Apple HealthKit and ResearchKit). Although control over who accesses one's data is a necessary pillar for privacy protection, it alone does not protect against more sophisticated types of privacy leaks, such as statistical inference attacks [Aranki and Bajcsy, 2015]. We will address the issue of privacy in mHealth and telemonitoring in Chapter 3. Finally, mHealth systems need to comply with local and federal regulations–such as HIPAA in the US. The systematic deployment of mechanisms that ensure or audit such compliance of mHealth systems remains an open challenge.

2.7 The Berkeley Telemonitoring Project

2.7.1 Introduction

In light of the lessons drawn from the CHF study in 2013 (Section 2.4), we took an endeavor to systematically study and address some of the challenges that we faced in designing the CHF telemonitoring system [Aranki et al., 2016b,a]. At the time, no similar open-source framework existed, to the best of our knowledge. This was the inception of the Berkeley Telemonitoring framework, from the Berkeley Telemonitoring project.[12] Even today, to the best of our knowledge,

[11] `https://clinicaltrials.gov/ct2/show/NCT02799407`

[12] The Berkeley Telemonitoring project: `https://telemonitoring.berkeley.edu`

there are no other similar general frameworks for telemonitoring research currently available for Android OS.

We first provide context to the goals of the Berkeley Telemonitoring project. Due to various factors including regulatory challenges, technical obstacles and the complexity of healthcare, mHealth remains largely limited to fitness and wellness applications [Aranki et al., 2017b]. On the other hand, mHealth has the potential to greatly benefit the treatment of chronic health conditions, including CHF, hypertension, diabetes and depression [Rickles et al., 2005; Paré et al., 2010]. Although such studies provided preliminary evidence to the effectiveness of mHealth and telemonitoring technologies to these conditions, most of them were only observational and limited to laboratory environment.

Moreover, telemonitoring technologies have the potential to reduce the prohibitive cost of long epidemiological studies, a thing that may expedite the realization of a full predictive healthcare model. This effort will require continued collaborations between technology and medical researchers, which may pose an extra challenge in the process. Hussain et al. [2015] argues that there is a lack in the involvement of healthcare professionals and researchers in the development of mHealth systems in general, and telemonitoring systems in particular. Therefore, it was evident in 2013 that there is a need for a framework that would allow easy development of health telemonitoring systems for *clinical research purposes*.

We chose to implement this framework for Android systems for the following reasons. First, Android OS has the highest share in the smartphone market, allowing our effort to have the highest impact. As of July 2017, Android market share was 85% (versus 14.7% for Apple iOS). Moreover, Android OS is an open-source platform; a fact that allows transparency in our efforts and allows independent contributions to our project. Finally, Hussain et al. [2015] state that the majority of the early mHealth apps targeted Apple iOS (since its release predated Android OS). Hussain et al. [2015] continue to assert that today, however, most of the mHealth research apps, either target Android OS alone or both Android OS and Apple iOS.

Identifying privacy as a key challenge in health telemonitoring, one of the core principles of the Berkeley Telemonitoring framework is privacy-aware design. In this section, we focus on the design aspects, and the features, of the Berkeley Telemonitoring framework. We detail the general privacy aspects of telemonitoring in Chapter 3. However, since privacy is part of the core design principles of the Berkeley Telemonitoring framework, we will also touch on some of the design decisions that were inspired by privacy considerations.

The Berkeley Telemonitoring framework provides libraries for the development of Android telemonitoring apps (client) as well as telemonitoring servers. The client can be any Android-enabled device, such as a smartphone, a tablet or a smartphone. For simplicity, we will assume that the telemonitoring client is a smartphone, for the remainder of this section. We note that the choice to support Android devices brings clinical research support to that platform, in a complementary way to that of Apple ResearchKit to Apple iOS.

The design objectives of the Berkeley Telemonitoring framework are as follows. First, smartphones have limited computational, connectivity and energy resources. As such, systems built on top of smartphones are more vulnerable to interruptions than systems that are built for personal computers. To be concrete, Android OS may elect to stop the running of an app in the background if it needs to free resources. As a result, telemonitoring apps (being background apps by nature) have to be designed with this in mind, and have to recover gracefully from faults without losing any collected data that were not yet submitted to the server at the time of the interruption. Therefore, the first objective in designing the Berkeley Telemonitoring framework is to relocate the responsibility of fault-tolerance from the app to the framework.

Second, because of the limitation of smartphones in their energy resources, telemonitoring apps should be careful not to consume too much power. Being careless with battery consumption may lead to poor adoption of telemonitoring systems because users may ultimately uninstall the telemonitoring apps from their smartphone [Aranki et al., 2016b]. Thus, the second design objective of the Berkeley Telemonitoring framework is to i) design the framework in an energy-preserving manner; and ii) design tools that enable telemonitoring apps to delegate computationally-demanding tasks to remote servers where energy consumption is a less pressing issue.

Third, we acknowledge that the main purpose of telemonitoring nodes is to collect health-related data. Because of that, telemonitoring systems become privacy-sensitive systems. It is therefore essential to design the Berkeley Telemonitoring framework in a privacy-preserving manner. To give a concrete example, assessing a patient's risk of clinical deterioration may require access to the data submitted by the rest of the monitored population (for relative comparison, for example). Granting the patient's smartphone access to population data, or statistics of them, will incur unwanted privacy leaks that needs to be prevented. Alternatively, by allowing the patient's telemonitoring app to delegate the risk assessment computation to the server may alleviate this risk. This is because the server already possesses access to the data submitted by the rest of the monitored population and can perform this task without needing to send statistical extracts of such data to the smartphone of the patient in question. We note, however, that this delegation of computation, in addition to data access and encryption controls and mechanisms, will not be sufficient to protect patients from more complex privacy attacks, such as statistical inference attacks. We address the topic of privacy in telemonitoring, in light of statistical inference attacks, in Chapter 3.

Fourth, on the same topic of collecting health-related data, we recognize that such collection is one of the main goals of telemonitoring systems. As a result, the Berkeley Telemonitoring framework needs to allow developers access to measured and estimated vital signs from i) internal smartphone sensors (such as EE estimation); ii) external health and fitness devices (such as blood pressure monitors); and iii) self-reported means through survey-like instruments (such as symptoms).

Finally, one of our goals in the Berkeley Telemonitoring project is to engage with healthcare researchers and professionals in designing mHealth systems. As such, the API of the Berkeley Telemonitoring framework has to be simple to use, such that non-technical developers are able to utilize it. In contrast, the Berkeley Telemonitoring framework needs to be modular and flexible in order to allow more seasoned software developers to extend its functionality. Therefore, the design of the Berkeley Telemonitoring framework needs to conform with proper software engineering design practices such as the Object-Oriented Programming (OOP) principles.

We summarize the design objectives of the Berkeley Telemonitoring framework in Table 2.5.

2.7.2 Framework Structure

There are three libraries shipped with the Berkeley Telemonitoring framework: i) *client library*; ii) *server library*; and iii) *core library*. The client library provides tools for telemonitoring apps, the server library provides tools for telemonitoring servers while the core library contains the common data structures and infrastructures that are shared by both other libraries. The breakdown of the Berkeley Telemonitoring framework is depicted in Figure 2.5.

All libraries are written in Java, for the following reasons. It is a natural decision to implement the client library in Java since it is geared towards Android systems. In order to create a uniform framework, we decided to implement the server library in Java as well, which allows us to i) share data structures between the libraries; and ii) utilize the data serialization mechanisms provided by

Identified Issue	**Design Objective**
Faults	The framework has to provides modules that are fault tolerant.
Battery consumption	Design the framework in an energy-mindful way and provide tools that enable developers to delegating computationally-intensive tasks to the server.
Privacy	The framework needs to be designed with privacy in mind. It should provide access controls, privacy-preserving data structures and protocols for privacy-preserving communication and data analysis.
Health-related data	The framework provide the necessary tools to developers to facilitate collecting health-related data by i) providing verified estimation algorithms relying on internal sensors; ii) providing the necessary APIs to access health data from external devices, wearables and sensors; and iii) providing mechanisms to collect self-reported data by means of survey-like instruments.
Ease of use	The framework needs to supply an easy-to-use API that hides the non-medically-related technical details.
Flexibility	In an effort to allow more advanced developers to extend its functionality, the framework's design needs to comply with software engineering design principles, such as the OOP principles.

Table 2.5: The design objectives and principles in designing the Berkeley Telemonitoring framework.

Java for client-server communication. As a result, the core library was also written in Java, since it is, intuitively speaking, the intersection of client and server tools and infrastructures.

To be concrete, the core library provides the necessary data structures for storage, fault tolerance, security and privacy, and surveys deployment. In contrast, the client library provides tools to i) connect to external wearables, sensors and devices for health data collection; ii) estimate health-related parameters and vital signs from internal sensors; iii) communicate with the server; iv) delegate computation; and v) render surveys on the smartphone screen. Finally, the server library provides tools to i) manage the collection and retention of data; ii) analyze data; iii) communicate with the smartphones; and iv) perform the requested delegated computation.

We now turn to describe the inner details of the Berkeley Telemonitoring framework. For more technical details about its implementation, we refer the reader to [Aranki et al., 2016a, 2017a].

Event-based Programming

We designed the Berkeley Telemonitoring framework to conform to the event-based programming paradigm. This paradigm allows software modules to request to be "updated" when certain events occur, similar to the concept of "hardware interrupts" from computer architecture. This translates to the telemonitoring app making requests to the framework, and immediately regaining CPU control. The framework will service that request and inform the telemonitoring app whenever the status of the request changes. The instances of updating the telemonitoring app about these status changes are referred to as *events*. The software constructs that implement the behavior of the app when events occur are called *listeners*.

Data Storage Paradigm

Given the fault-prone nature of mobile systems, such as Android OS killing background services and apps in order to free resources, we designed resilient, fault-tolerant, data storage structures in the Berkeley Telemonitoring framework. That is, when Android OS elects to kill the telemonitoring

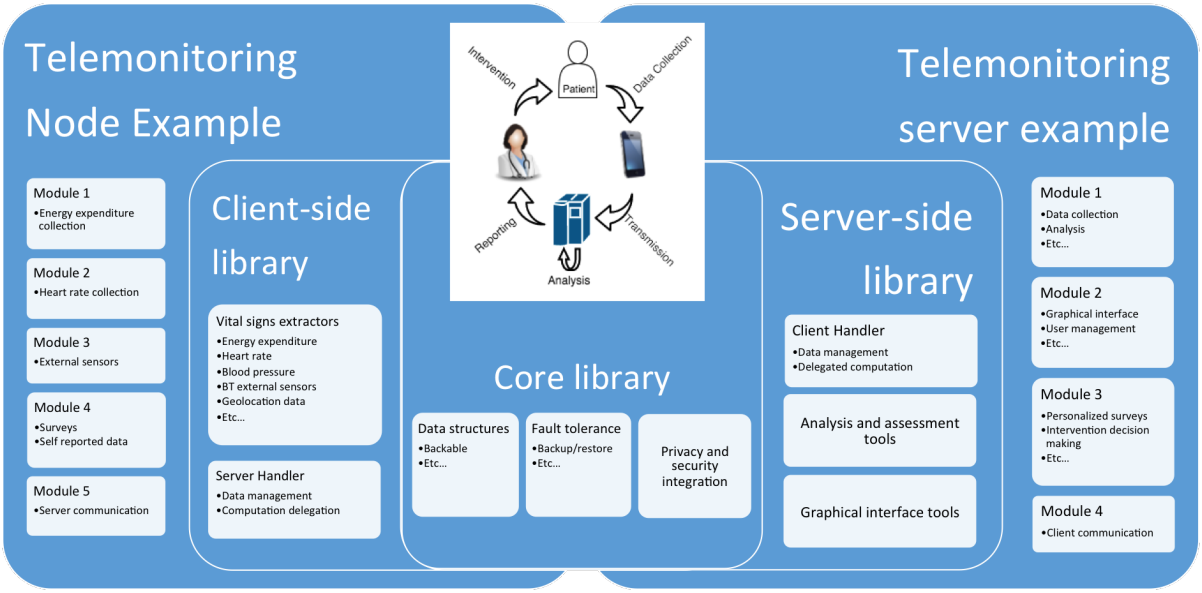


Figure 2.5: The architectural breakdown of the Berkeley Telemonitoring framework.

app, instead of dismissing the collected data that were not yet submitted to the server, these structures make sure that the data are retained and restored whenever the telemonitoring app resumes to operate. This is done through retaining a copy of the data on non-volatile storage and updating it as new datapoints are collected. In regular scenarios, the telemonitoring app itself has to design such fault tolerance, which can be complex and requires a careful implementation.

Instead, the Berkeley Telemonitoring framework provides this fault-tolerance in the library level so that telemonitoring apps don't have to implement it themselves. The framework, therefore, provides data-structures that are immediately backed up on resilient storage in any event of change in their content. These data structures are called *backables*. The mechanism that performs this

Algorithm 2.1 A code snippet to create backables for heart-beat rate and cadence data, and registering them with a backup cabinet.

```
1  // Start a backup cabinet
2  BackableEncapsulatorBackupCabinet cabinet = new
       BackableEncapsulatorBackupCabinet("/path/to/storage");
3  // Create an identifier for the heart-beat rate backable
4  StringIdentifier hrDataID = new StringIdentifier("heart rate data");
5  // Create the encapsulator for heart-beat rate data
6  BackableEncapsulator<HeartRateContainer> hrEnc = new
       BackableEncapsulator<HeartRateContainer>(hrDataID);
7  // Create an identifier for the cadence backable
8  StringIdentifier cadenceDataID = new StringIdentifier("cadence data");
9  // Create the encapsulator for cadence data
10 BackableEncapsulator<CadenceContainer> cadenceEnc = new
       BackableEncapsulator<CadenceContainer>(cadenceDataID);
11 // Register the backables with the cabinet
12 cabinet.registerBackable(hrEnc);
13 cabinet.registerBackable(cadenceEnc);
```

Algorithm 2.2 A code snippet to create a backup cabinet located in the private app storage area.

```
1 // Start a backup cabinet in the private app area
2 BackableEncapsulatorBackupCabinet cabinet = new
      BackableEncapsulatorBackupCabinet(new
      File(getApplicationContext().getFilesDir(), "/my_backup"));
```

bookkeeping is called a *backup cabinet*. A backup cabinet is assigned a storage location for its operation. Once a backable data structure registers to be backed up by a backup cabinet, the cabinet checks whether it has an old copy of that backable in storage. If the answer is positive, the backup cabinet restores the old copy into the backable and the telemonitoring app resumes operation. From this point onwards, the backup cabinet updates its storage to reflect the changes occurring in the backable. Algorithm 2.1 lists an example code snippet that creates two backables, one for heart-beat rate data and another for cadence data, and registers them with a backup cabinet that stores the data at `"/path/to/storage"`.[13]

Privacy and Security

Multiple measures are designed in place to ensure that the Berkeley Telemonitoring framework is privacy-aware and has adequate security standards. First, to ensure that the stored backups of the data stored in backables are secure from unauthorized access or tampering, backup cabinets utilize the storage options provided by the Android OS to store the data in an area that is accessible only by the same app that generated them. This can be achieved by initializing the backup cabinet as listed in Algorithm 2.2. Furthermore, we incorporate privacy-preserving mechanisms that are implemented on top of backables so that telemonitoring systems developers benefit from them without the burden of implementing them by themselves. In particular, some of these mechanisms are designed to protect the telemonitoring users from statistical inference attacks. These mechanisms are based on Private Disclosure of Information (PDI), which is presented in great detail in Section 3.4.

Bluetooth and BLE

As stated in the design objectives of the Berkeley Telemonitoring framework, on the primary objectives of telemonitoring is collecting health-related data. As such, the Berkeley Telemonitoring framework provides tools to access and collect such data from external sensors, wearblaes and devices. These tools are implemented on top of Bluetooth and Bluetooth Low Energy (BLE), as most of the consumer wearables and devices are BLE enabled.

In order to be consistent with the event-based programming paradigm that the framework adopts, the Berkeley Telemonitoring framework extended the Android Bluetooth and BLE capabilities in a manner that unifies their behavior [Azar et al., 2015]. The original Bluetooth stack in Android follows a polling, busy-wait paradigm. This means that the telemonitoring app would have to actively and regularly check for updates in the status of the connections it maintains. For example, the app would have to manually check to identify events of establishing a new connection, receiving a message, etc. BLE, on the other hand, is event-based by design; and Android's stack

[13] Cadence is defined as the number of steps taken per minute.

Algorithm 2.3 An example listener for heart-beat rate and blood pressure data.

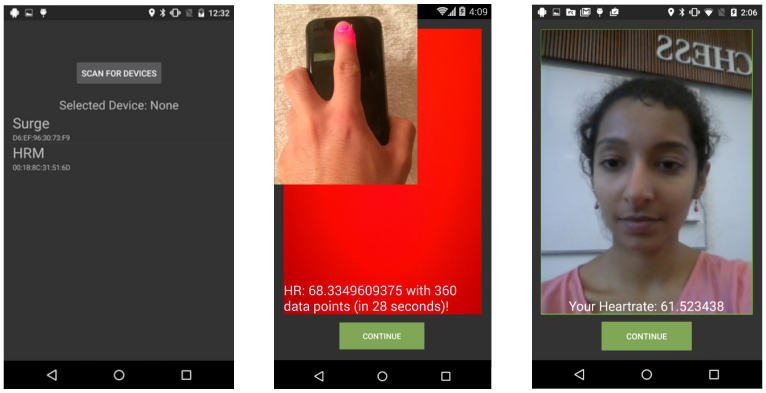
```
1  public class MyHealthDataListener implements
2          BluetoothHeartRateListenerInterface,
3          BluetoothBloodPressureListenerInterface {
4
5      public void heartRateDataPointReceived(BluetoothHealthDevice device,
6          HeartRateContainer datapoint) {
7          // code to run when a heart rate datapoint is received (because this
               implements BluetoothHeartRateListenerInterface)
8      }
9
10     public void bloodPressureDataReceived(BluetoothHealthDevice device,
11         BloodPressureContainer datapoint) {
12         // code to run when a blood pressure datapoint is received (because
               this implements BluetoothBloodPressureListenerInterface)
13     }
14
15     public void updateConnectionState(BluetoothHealthDevice device,
16         boolean connected) {
17         // code to run when the connection state changes
18     }
19 }
```

for it follows the same paradigm. As a result of this lack of uniformity, the Berkeley Telemonitoring framework extended both stacks in an effort to unify them both under the event-based paradigm. With the extended stack, telemonitoring apps place Bluetooth or BLE requests with the framework and continue to perform other tasks. The framework will invoke the appropriate listener(s) in the app whenever an event pertaining to these requests occurs, as discussed earlier. It is important to mention that these extended stacks also supply embedded fault-tolerance mechanisms that deal with interruptions in the connections, without the involvement of the app layer itself. Figure 2.6a depicts a screenshot of an app using the Bluetooth stack to scan for nearby devices.

Using the extended Bluetooth and BLE stacks described above, the Berkeley Telemonitoring framework provides native access to devices that are compliant with the PHD standard [EMB/11073, 2012]. In practice, the app implements the desired listener(s) for the different types of health data. The app indicates what data it wishes to handle in that listener by stating which

Algorithm 2.4 A standard snippet for communicating with ISO/IEEE 11073 Personal Health Device (PHD) enabled devices.

```
1 // Initialize the health device, you don't need to specify the type, make or
      model of the device
2 BluetoothHealthDevice mHealthDevice = new BluetoothHealthDevice(bleDevice);
3 // Register any health data listeners
4 mHealthDevice.addHealthListener(new MyHealthDataListener());
5 // Request to connect and start receiving data
6 mHealthDevice.connect();
```



(a) (b) (c)

Figure 2.6: (a) A screenshot of Bluetooth/BLE stack scanning for nearby devices; (b) a demonstration of the finger-based heart rate estimator (top: an external photo of the setting, bottom: the interface in the telemonitoring app); and (c) a demonstration of the face-based heart rate estimator.

interfaces that listener implements. Algorithm 2.3 lists a snippet sample for an implementation of a listener (`MyHealthDataListener`) that handles heart-beat rate data as well as blood pressure data. At this point, in order to connect to a PHD-enabled device, the telemonitoring app needs to only define an object representing the PHD device, register any listener(s) to it to indicate the behavior when these devices collect new datapoints, and request to connect to it. After this point, the framework will proceed to connect to the device and start extracting data. Algorithm 2.4 lists a snippet sample for connecting to a PHD-enabled device and registering the listener `MyHealthDataListener` with it. These snippets demonstrate the ease-of-use of the framework's API in communicating with PHD-enabled devices from telemonitoring apps, without the need to implement complex and detailed protocols, and deal with fault tolerance.

The Berkeley Telemonitoring framework currently supports the following PHD standards: i) IEEE/ISO 11073-10404: pulse oximeters; ii) IEEE/ISO 11073-10407: blood pressure monitors; iii) IEEE/ISO 11073-10408: thermometers; iv) IEEE/ISO 11073-10415: weighing scales; and v) IEEE/ISO 11073-10417: glucose meters [EMB/11073, 2012].

Estimators and Extractors

Another potential source of health-related data is the internal smartphone sensors. This can be achieved by estimating health-related markers from sensors such as accelerometers. Therefore, the Berkeley Telemonitoring framework provides algorithms that can i) extract raw sensory data from the smartphone, which we call *extractors*; and ii) estimate health and fitness related parameters from internal sensors, which we call *estimators*. Backables are used by extractors and estimators to store the data they produce. Extractors and estimators start producing datapoints whenever the telemonitoring app requests that they start and continue to do so in the background, filling the provided backable, until the app requests them to stop.

At the time of writing this dissertation, the Berkeley Telemonitoring framework supported the following extractors and estimators: i) GPS extractor, which allows the telemonitoring app to regularly learn the smartphone GPS location; ii) call status extractor, which allows the telemonitoring app to regularly learn whether a phone call is in session or not; iii) battery extractor, which allows

Algorithm 2.5 A code snippet that uses a cadence estimator, requesting an update every 3 seconds.

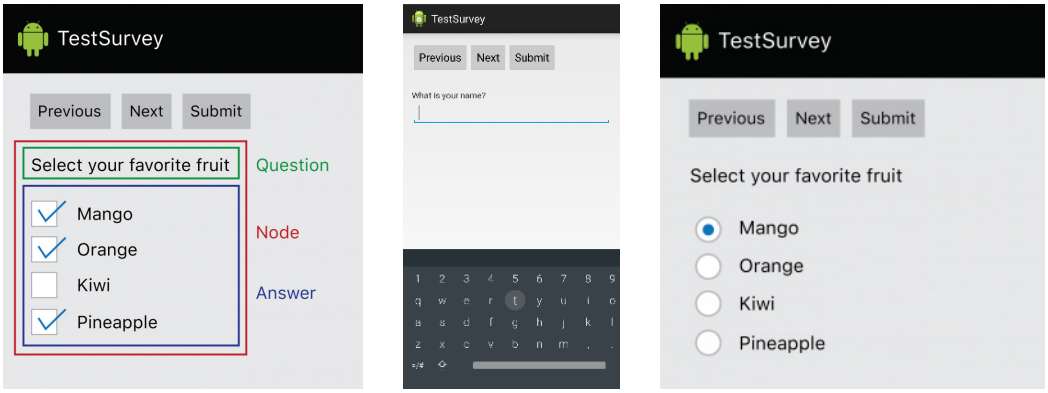
```
int msPeriod = 3 * 1000; // Three Seconds
// Initialize the cadence backable
StringIdentifier cadenceId = new StringIdentifier("Cadence data");
BackableEncapsulator<TimeZonedTimestampedObject<CadenceContainer>> cadenceData
    = new BackableEncapsulator<>(cadenceId);
// Initialize the cadence estimator with 3 seconds period of updates
CadenceEstimator cadenceEstimator = new CadenceEstimator(cadenceData, msPeriod,
    this.getApplicationContext());
// Start the estimator
cadenceEstimator.start();
```

the telemonitoring app to regularly learn the smartphone battery level and whether the phone is charging or not; iv) screen light extractor, which allows the telemonitoring app to regularly learn whether the smartphone screen is on or off; v) EE estimator, which allows the telemonitoring app to get regular estimates of EE from accelerometer data [Chen and Sun, 1997; Donaire-Gonzalez et al., 2013]; vi) heart-beat rate estimator from a finger video, which allows the telemonitoring app to get real-time estimates of the heart-beat rate from a video feed of the subject's index finger that is placed over the camera [Azar et al., 2015]. Figure 2.6b depicts an example of using this estimator; vii) heart rate estimator from a face video, which allows the telemonitoring app to get real-time estimates of the heart-beat rate from a video feed of the subject's face [Poh et al., 2011; Azar et al., 2015]. Figure 2.6c depicts an example of using this estimator; viii) accelerometer-based speed estimator, which allows the telemonitoring app to get regular estimates of the smartphone's travel speed from accelerometer data [Park et al., 2012; Azar et al., 2015]; ix) GPS-based speed estimator, which allows the telemonitoring app to get regular estimates of the smartphone's travel speed from GPS data [Aranki et al., 2017a]; x) GPS-based distance estimator, which allows the telemonitoring app to get regular estimates of the smartphone's traveled distance from GPS data [Aranki et al., 2017a]; xi) cadence estimator, which allows the telemonitoring app to get regular estimates of the subject's cadence (steps per minute) from accelerometer data [Mladenov and Mock, 2009; Asuncion et al., 2016]; Algorithm 2.5 lists an example code snippet that obtains and starts a cadence estimator.

Surveys

The last source of health-related data in our design objectives is self-reported data. The Berkeley Telemonitoring framework provides tools to deploy surveys in telemonitoring apps. These tools include the ability to codify surveys, gather the subject responses to the surveys and render surveys on the smartphone. We represent a *survey* as a list of *survey nodes*. Each survey node consists of a pair of a *question* and an *answer*. The survey node question represents the question portion, which can be a text question, a picture question, etc. Similarly, the survey node answer represents the answer portion, which can be a text answer, a check list answer, a radio list answer, etc. Figure 2.7a visually labels the different components of a survey, whereas Figures 2.7b and 2.7c depict examples of different combinations of question/answer survey nodes.

In order to simplify the deployment of surveys, we provide an easy-to-use API that can be used to render the survey objects on the smartphone screen. We call these constructs *survey renderers*.



(a) (b) (c)

Figure 2.7: (a) The different components of the survey; (b) an example of a text question/answer survey node; and (c) an example of a text question and radio answer survey node.

Survey renderers convert the different components of a survey into Android fragments that can be placed anywhere. This design a) is modular and allows telemonitoring apps to deploy surveys with any combination of question and answer types; b) simplifies the deployment of these surveys by providing automatic rendering capabilities; and c) allows developers to implement new types of questions and/or answers for further support. Algorithm 2.6 lists an example code snippet that renders the survey nodes in a given survey.

Client-Server Communication

The Berkeley Telemonitoring framework also provides tools to facilitate the communication between the telemonitoring app and server. Traditionally, the burden of implementing such a communication protocol falls on the developer of the telemonitoring system, with all the fault tolerance that it entails. In this module, we intend to elevate that responsibility to the framework level. To do so, we break down the communication to units of jobs. In particular, the Berkeley Telemonitoring framework defines two central types of jobs.

Data Jobs Data jobs encapsulate the app's intent to transmit data to the server. These jobs can be used to submit the collected health-related data from the telemonitoring app, such as vital signs and surveys.

Algorithm 2.6 A code snippet to render all survey nodes in a given survey.

```
1 // Given a survey object "survey," create a SurveyRenderer object from it
2 SurveyRenderer surveyRenderer = new SurveyRenderer(survey, new
      SurveyConverter());
3 // Iterate over the nodes
4 for (SurveyNode<?,?> sn: surveyRenderer) {
5     // Render the current survey node
6     Fragment currFragment = ((SurveyNodeRenderer<?,?>) sn).render();
7 }
```

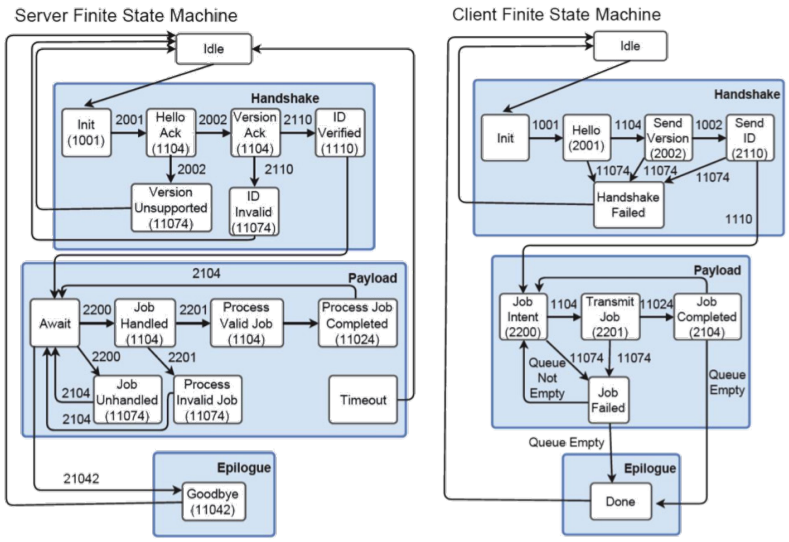


Figure 2.8: The finite state machine description of the TI protocol.

Request Jobs Request jobs encapsulate the app's intent to request that some computation be carried on the server. As outlined in our discussion on the design objectives of the framework in Section 2.7.1, such delegation of computation have the following benefits:

1. Battery consumption: computationally-intensive tasks may incur a high load on the smartphone's batter, which may hinder adoption. Instead, requesting that such computational tasks be carried on the server alleviates this problem; and

2. Privacy: certain jobs, including some predictive analysis of the monitored subject's clinical deterioration risk may require access to data pertaining to other individuals than the subject in question. Allowing the subject's phone to get access to such data, or statistics of them,

Algorithm 2.7 A code snippet to start a telemonitoring server that handles one job using the `MyJobListener` job listener.

```
// server listens or port 9999
int port = 9999;
// the allowed TLS ciphers
String[] ciphers = {"TLS_RSA_WITH_AES_128_CBC_SHA"};
// the TLS keystore path and password
String ksPath = "/path/to/keystore";
String ksPassword = "password";
// Obtain a TI protocol object
TIProtocol tiProtocol = new TIProtocol(ciphers, port, ksPassword, ksPath);
// Obtain a client handler, the module for communication fault tolerance
ClientHandler ch = new ClientHandler(tiProtocol);
// Register listeners that handle the jobs from the client
MyJobListener listener = new MyJobListener();
ch.addRespondingJobListener(listener);
// Start the server
ch.start();
```

may cause a privacy leak that needs to be avoided [Dwork, 2006; Aranki et al., 2017b]. Instead, requesting that this analysis be performed by the server, which already has access to such privacy-sensitive data (and is authorized to have such access), is a safer approach from a privacy point of view.

The Berkeley Telemonitoring framework provides the necessary definitions to allow developers to implement specific protocols that can communicate in the units of these jobs. This breakdown allows the framework to implement fault-tolerant mechanisms on top of these specific protocols, without knowing their implementation details. We refer to these specifications as the communication *meta-protocol*, which is the set of requirements for any communication protocol to benefit from the fault-tolerance tools in the framework. Any communication protocol satisfying these specifications is considered a valid telemonitoring communication protocol.

This design allows developers to implement their own protocols or adapt their existing protocols to benefit from the fault-tolerance mechanisms that are supplied by the framework [Aranki et al., 2016a]. This addresses the design objective of flexibility, as discussed in Section 2.7.1. In addition to this, and in order to satisfy our design objective of ease-of-use, the Berkeley Telemonitoring framework also provides a ready-to-use telemonitoring communication protocol, *Tele-interfacing (TI)* protocol, which is implemented on top of Transport Layer Security (TLS), or its predecessor, SSL, over Transmission Control Protocol (TCP). The finite state machine description of the TI protocol is depicted in Figure 2.8. Algorithm 2.7 lists a code snippet that demonstrates the simplicity of starting a server using the TI protocol. Algorithm 2.8 lists a code snippet that demonstrates the simplicity of communicating with the telemonitoring server from the app, using `ServerHandler`.

Algorithm 2.8 A code snippet to obtain a `ServerHandler` for the telemonitoring app to communicate with the server in a fault tolerant way; and an example of sending a data job.

```
1  // My user identifier
2  BasicUserIdentifier uid = new BasicUserIdentifier("My user ID");
3  // My version identifier
4  VersionIdentifier vid = new VersionIdentifier(1,0,0); // Version identifier
5  // Where do the servers physically reside?
6  TIServerAddress[] serverAddresses = {
7      new TIServerAddress("server.address.one", 9999),
8      new TIServerAddress("server.address.two", 9999)
9      // etc...
10 };
11 // Define a server identifier, using the matching TLS keystore
12 TIServerIdentifier si = new TIServerIdentifier(keyStore, "keyStorePassword",
       serverAddresses);
13 // Obtain a TI protocol object
14 TIProtocol protocol = new TIProtocol(si, vid, uid);
15 // Obtain a server handler, for fault-tolerant communication handling
16 ServerHandler sh = new ServerHandler(context, protocol, connectionInterval,
       pathForBackupCabinet); // Create a server handler
17 // From this point, we can use this server handler to send request and data
       jobs.
18 // Request that some data be sent to the server, all events pertaining to this
       job will be delivered to jobListener
19 sh.sendData(dataID, jobListener);
```

Server Tools

Finally, the Berkeley Telemonitoring framework provides tools for the server side of any telemonitoring system. The first major category of tools that the framework supplies on the server side is job handling capabilities. These tools allow servers to handle jobs (data jobs, request jobs, etc.). The main construct in this category is a *job listener*. There are two types of job listeners, a *responding* job listener and a *non-responding* job listener. The responding job listeners indicate that once they process the job, they wish to provide a reply to the client (e.g., return value). The non-responding job listeners simply process the job without having anything to reply back to the client (and so the `ClientHandler` does not wait for those before it replies to the client).

The second category of server tools provided by the framework is data curation capabilities. The basic building block of this category is called a *table modifier*. For each type of health-related data that the framework natively supports, it pairs it with a SQL table structure that can store it and a corresponding table modifier that can be used to alter the contents of that table. This allows

Algorithm 2.9 A code snippet that implements a job listener `MyJobListener` (see Algorithm 2.7) that stores EE data in a Structured Query Language (SQL) database, using a `EETableModifier`.

```
1  public class MyJobListener extends AbstractNonrespondingClientJobListener {
2      // Checks, per job, whether this handler wants to handle it
3      @Override
4      public boolean isJobHandled(AbstractServerJob<?, ?> job) {
5          if (job instanceof DataJob) {
6              DataJob dataJob = (DataJob) job;
7              return dataJob.getJobIdentifier().getStringId().equals(eeJobName);
8          }
9          return false;
10     }
11     // The method that gets invoked to handle the job
12     @Override
13     public void processJob(UserIdentifierInterface userID, AbstractServerJob<?,
           ?> job) {
14         // Get a table modifier
15         EETableModifier eeMod = new EETableModifier(eeTableName);
16         // Cast to data job
17         DataJob dataJob = (DataJob) job;
18         // Expecting a job with encapsulator data of type
               TimeZonedTimestampedObject<EEDataContainer>
19         for (DataSerialPair<? extends Serializable> dataPoint :
               dataJob.getData()) {
20             // Cast to the expected data type
21             TimeZonedTimestampedObject<EEDataContainer> dp =
                   (TimeZonedTimestampedObject<EEDataContainer>)
                   dataPoint.getDatapoint();
22             // Store the data point in the SQL database.
23             eeMod.insertRecord(dp.getObject(), dp.getTimeZonedTimestamp(),
                   userID);
24         }
25     }
26 }
```

us to provide mechanisms to store, retrieve, edit and/or delete records from these tables without exposing the developer to SQL queries. Algorithm 2.9 lists a code snippet that implements the non-responding job listener `MyJobListener` (see Algorithm 2.7) to store the EE data contained in a data job that contains EE data.

The third category of server tools provided by the framework is data analysis tools. These include statistical and machine learning tools that may be used to construct models from the collected data. We are currently working on expanding the base of algorithms in this category.

2.8 RunningCoach Study

2.8.1 The Premise

In an effort to test the feasibility of the Berkeley Telemonitoring framework, we designed a study for telemonitoring of long-distance runners. The aim of the monitoring is to help the runners optimize their cadence in order to minimize their risk of injury.[14] The study was approved by the Institutional Review Board at UC Berkeley. The premise for this intervention is as follows. Connections have been demonstrated between running at a proper cadence and i) the reduction of impact forces on joints [Heiderscheit et al., 2011]; ii) the reduction of fatigue and muscle soreness [Rowlands et al., 2001]; and iii) the increase of efficiency of oxygen use [Hamill et al., 1995].

As such, we designed RunningCoach, a telemonitoring system that collects running-related data from the runners, during their runs, and provides periodic interventions, based on such data. We will briefly describe the design of the system and the study, but we refer the reader to [Aranki et al., 2017a] for more details. RunningCoach is implemented using the Berkeley Telemonitoring framework, including both the telemonitoring smartphone app and the telemonitoring server. In the next section, we briefly describe the design of the telemonitoring system.

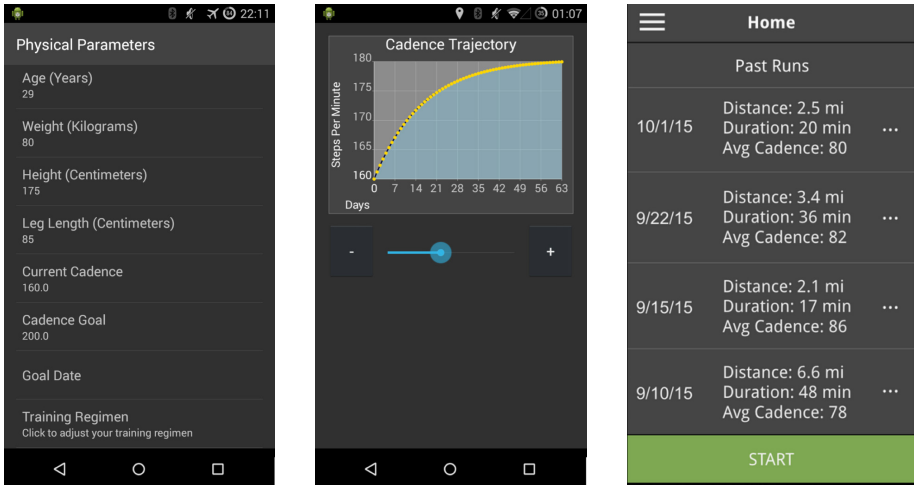
2.8.2 System and Study Design

As mentioned earlier, RunningCoach was built using the Berkeley Telemonitoring framework. This allows us to take advantage of all the functionalities describes in Section 2.7. Moreover, this will allow us to test the usability of the framework with a real research application. RunningCoach aims to help long-distance runners achieve a target cadence within a provided time frame. In order to achieve that, the app collects data about the physical parameters of the runner, as depicted in Figure 2.9a. The runner sets her or his initial cadence as part of the inquiry, and sets the desired target cadence to be achieved by a set date. As a result, the app produces a training regimen for the runner that follows the following exponentially decaying form

$$C(d) = \frac{C_N \cdot e^{\alpha N} - C_0}{e^{\alpha N} - 1} - \frac{C_0 - C_N}{e^{-\alpha N} - 1} \cdot e^{-\alpha d}$$

where $C(d)$ is the cadence training regimen as a function of day d; C_0 and C_N are the initial and target cadence values, respectively (steps/minute); α is the decay factor in the training regimen; and N is the duration of the training regimen (in days). Note that this model is the solution to the function $C(d) = A + B \cdot e^{-\alpha d}$ with initial conditions $C(0) = C_0$ and $C(N) = C_N$. We also note that when α approaches 0, the training regimen approaches a linear curve [Aranki et al., 2017a]. An example of such a training regimen is depicted in Figure 2.9b. The rationale behind a gradual

[14] Cadence is defined as the number of steps taken per minute.



(a) (b) (c)

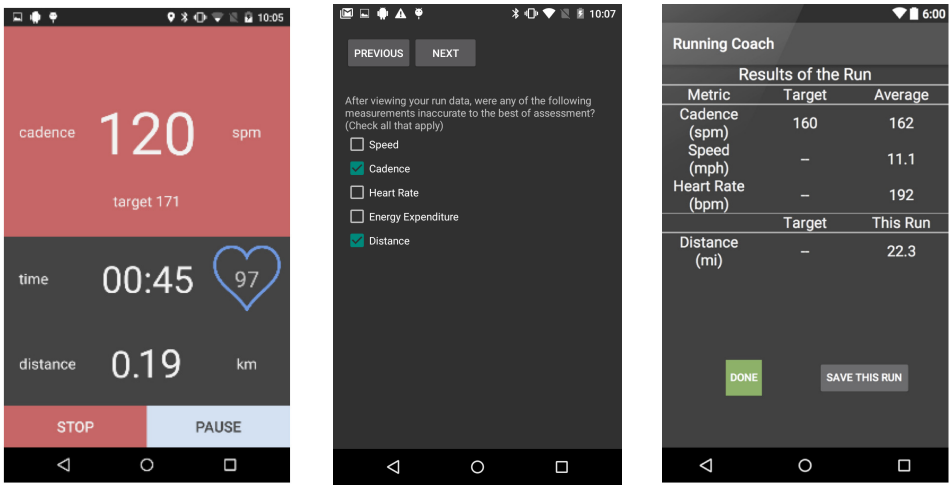
Figure 2.9: Screenshots from the RunningCoach app: (a) the runner's physical parameters screen; (b) an example cadence training regimen; and (c) the app's home screen.

convergence to the target cadence is to minimize any risk that may occur due to a sudden change in the runner's routine.

The home screen of RunningCoach, depicted in Figure 2.9c, displays a list of past runs and statistics about them, such as the length and duration. From the home screen, the runner can access the app settings, allowing her or him to select which data she or he wishes to share with the study team, and which she or he wishes to keep private. Also from the home screen, the runner can start a run, by clicking on the **START** button. After that, the app asks the runner whether she or he wishes to take two heart-beat rate measurements, one from a video of her or his index finger (as discussed earlier and depicted in Figure 2.6b), and another one from a video of her or his face (as discussed earlier and depicted in Figure 2.6c). The runner may choose to take or skip either or both of these measurements.

During the run, the app is designed to collect regular updates about the runner's performance and health status, complying with her or his privacy settings as discussed earlier. The collected data variables include i) GPS location; ii) cadence; iii) EE; iv) heart-beat rate (through a provided PHD enabled chest-strap heart monitor); v) speed; vi) distance covered; vii) battery information; and viii) screen light information. The screen that is displayed on the screen during the run is depicted in Figure 2.10a. It is important to mention that the monitoring actually runs in the background, meaning that it will continue to run even if the runner "minimized" the RunningCoach app or even locked her or his phone. During the run, RunningCoach provides intervention to the runner whenever her or his cadence falls far from the target cadence for the day, according to the training regimen. RunningCoach will only provide the intervention if the runner is not running within 10% of the target cadence for some period of time (20 to 30 seconds, depending on the the settings). The intervention is delivered through auditory and haptic means (vibration).

The run continues until the runner terminates it by clicking on the **STOP** button. Once this event occurs, the app prompts the runner to take two other measurements of heart rate using the video-based estimation algorithms as described earlier. Also in this instance, the runner may choose to take or skip either or both measurements. After these measurements, the app prompts the runner with a post-run survey that inquires about the usability of the app. An example of



(a) (b) (c)

Figure 2.10: Screenshots from the RunningCoach app: (a) the screen that is displayed during the run; (b) a sample post-run survey question; and (c) a sample post-run summary.

a survey screen is depicted in Figure 2.10b. The runner may choose to skip any question in the survey. Once the survey part is completed, the app displays the summary statistics of the run, as depicted in Figure 2.10c.

The telemonitoring server is designed to receive all the collected data from the RunningCoach app, store them, and provide visualization for them. Figure 2.11 depicts various screenshots of the collected data from the server visualizations.

The design of RunningCoach was easy and bump-free, compared to the design of the CHF telemonitoring app from the CHF study described in Section 2.4. We believe that this is due, by large, to the existence of and wide range of tools supplied by the Berkeley Telemonitoring framework. During the design of RunningCoach, our full attention and focus was on the study aspects of the system, and less on the systems considerations, such as fault-tolerance, communication protocols and estimation algorithms implementation. This was only possible because these aspects are all provided natively by the Berkeley Telemonitoring framework. We will report on the privacy and acceptability of RunningCoach as a result of this study in Section 3.2.3.

2.9 Summary and Discussion

In this chapter we addressed the question of streamlining and standardizing health-related data collection. Streamlining health-related data collection has the potential to drop the prohibitive cost of long-term epidemiological studies, which are necessary for achieving the predictive healthcare model. We identified a system architecture that can both serve as a standardized means for health-related data collection, and as a system that can implement mHealth-based predictive medicine. This architecture is telemonitoring.

We then presented a study of telemonitoring in patients with CHF. We drew conclusions from the study regarding the design of telemonitoring systems, and identified the need for a general and standardized framework for telemonitoring. As a result, the Berkeley Telemonitoring project was born and the work to build the Berkeley Telemonitoring framework began. We further described the inner workings of the Berkeley Telemonitoring framework, and provided examples of its ease-

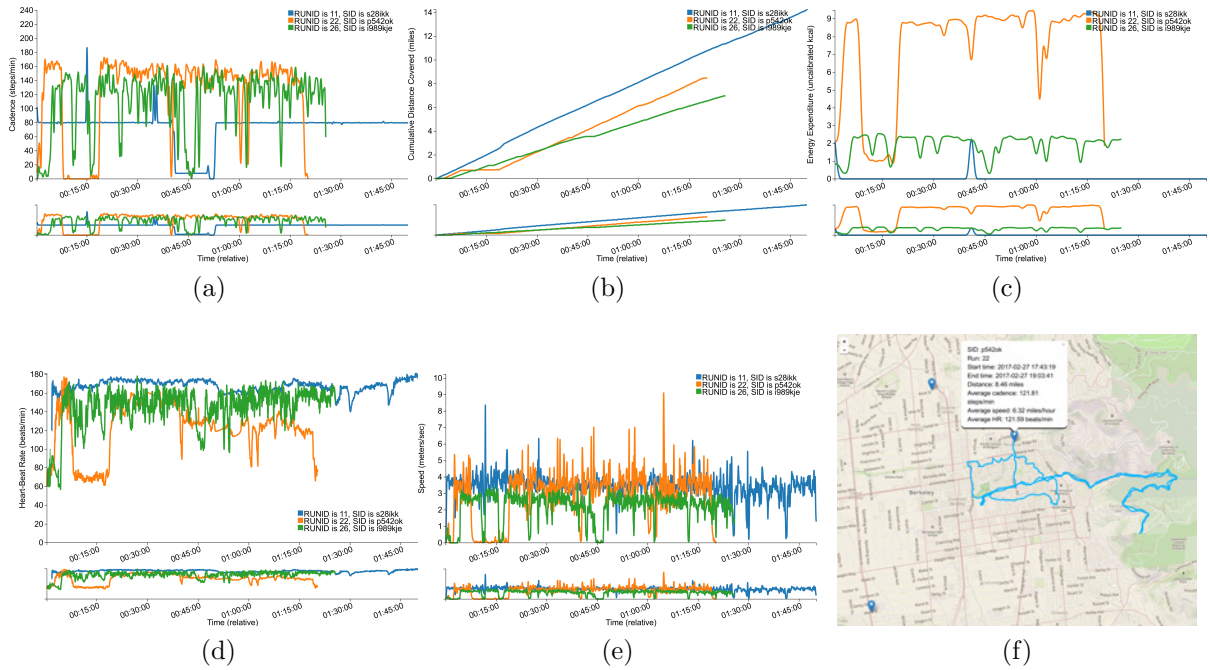


Figure 2.11: Sample plots from the server dashboard: (a) cadence plot; (b) distance-covered plot; (c) EE plot; (d) heart-beat rate plot; (e) speed plot; and (f) GPS plot showing the path of run 22.

of-use and its flexibility. As an example of its ease of use, we presented a study, RunningCoach, that was designed on top of the Berkeley Telemonitoring framework. We described the study and system design of RunningCoach and presented examples of its features.

Even though the work on the Berkeley Telemonitoring framework preceded the release of similar commercial frameworks, we surveyed the market for other solutions in order to provide a complete-picture context for our work. We discussed the notion of medical intervention, and discussed the challenges in the fields of mHealth and telemonitoring. Some of these challenges were addressed in the Berkeley Telemonitoring framework, but some others remain open. More research is necessary to resolve these challenges and get us closer to a fully working predictive healthcare model. We address two of these challenges in the next two chapters of this dissertation. One challenge that spans beyond predictive medicine is assessing the reliability of scientific findings from scientific literature. We address this challenge and present a mathematical treatement for it in Chapter 4. But first, let us discuss the challenge of privacy in telemonitoring in Chapter 3.

Acknowledgments

Many people have directly and indirectly helped in preparing the manuscript of this chapter, both their support and feedback are greatly appreciated. First and foremost, I would like to thank my dissertation committee members, Professors Ruzena Bajcsy, John Canny and Deirdre Mulligan for their contributions, support, feedback and ideas, which greatly improved the quality of this chapter. Gregorij Kurillo was instrumental in many aspects of this chapter, he is a true pillar on which much of this chapter stands. I would like to thank David M. Liebovitz, MD for his continued support in telemonitoring and mHealth research. The CHF study, which taught me much of what

I know about telemonitoring, would not have been possible without the tireless work of Ariane Garrett, Mita Goel, MD/MPH, Enid Montague, PhD, Robert A. Gordon, MD, Daniel Schimmel, MD, Chantal M. Mendes and the entire staff at Northwestern Medical Faculty Foundation; for that I am eternally grateful to all of them. I would also like to thank Heather M. Patterson, Martin French and Helen Nissenbaum for their contribution in the design of the privacy and acceptability survey as well as other privacy-related aspects of the CHF study. The Berkeley Telemonitoring Project would not have been possible without the dedication of every member of its team–their contributions are greatly appreciated and valued.[15] Explicitly, I would like to thank Posu Yan, Arjun Chopra, Eugene Song, Adarsh Mani, Phillip Azar, Jochem van Gaalen, Quan Peng, Priyanka Nigam, Maya P. Reddy, Sneha Sankavaram, Qiyin Wu, Uma Balakrishnan, Hannah Sarver, Lucas Serven, Carlos Asuncion, Kaidi (Kate) Du, Caitlin Gruis, Gao Xian Peh, Yu (Sean) Xiao and Joany Gao.

This work was supported in part by TRUST, Team for Research in Ubiquitous Secure Technology, which receives funding support for the National Science Foundation (NSF award number CCF-0424422). This manuscript was made possible by Grant Number HHS 90TR0003/01. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the United States Department of Health and Human Services. This work was supported in part by the Center for Long-Term Cybersecurity (CLTC) at UC Berkeley. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the CLTC.

Any errors or mistakes that made it to the final version of this chapter, including typographical ones, are solely my responsibility, not that of any person or entity mentioned above.

Bibliography

Alshurafa, N., Eastwood, J.-A., Nyamathi, S., Liu, J. J., Xu, W., Ghasemzadeh, H., Pourhomayoun, M., and Sarrafzadeh, M. Improving Compliance in Remote Healthcare Systems Through Smartphone Battery Optimization. *Biomedical and Health Informatics, IEEE Journal of*, vol. 19(1):pp. 57–63, Jan 2015. ISSN 2168-2194. doi:10.1109/JBHI.2014.2329712.

Apple Inc. ResearchKit Programming Guide - Creating Surveys at http://researchkit.org/docs/docs/Survey/CreatingSurveys.html. 2016. Accessed: 09/10/2016.

AppleInsider Staff. Over 10K participants sign up for Stanford medical trial after ResearchKit debut at http://appleinsider.com/articles/15/03/11/over-10k-participants-sign-up-for-stanford-medical-trial-after-researchkit-debut. 2015.

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D., Balakrishnan, U., Sarver, H., Serven, L., Asuncion, C., Du, K., Gruis, C., Peh, G. X., Xiao, Y., and Bajcsy, R. RunningCoach – Cadence Training System for Long-Distance Runners. In *2017 Health-i-Coach – Intelligent Technologies for Coaching in Health*. May 2017a.

Aranki, D., Kurillo, G., and Bajcsy, R. Smartphone Based Real-Time Health Monitoring and Intervention. In Khan, S. U., Zomaya, A. Y., and Abbas, A. (eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, chap. ?? Springer, 2017b. In press.

[15] `https://telemonitoring.berkeley.edu/team/`

Aranki, D., Kurillo, G., Mani, A., Azar, P., van Gaalen, J., Peng, Q., Nigam, P., Reddy, M. P., Sankavaram, S., Wu, Q., and Bajcsy, R. A telemonitoring framework for android devices. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 282–291. IEEE, 2016a.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-time Tele-monitoring of Patients with Chronic Heart-Failure Using a Smartphone: Lessons Learned. *IEEE Transactions on Affective Computing*, 2016b. doi:10.1109/taffc.2016.2554118. URL `http://dx.doi.org/10.1109/TAFFC.2016.2554118`.

Asuncion, C., Balakrishnan, U., Sarver, H., Serven, L., and Song, E. *A Telemonitoring Solution to Long-Distance Running Coaching*. Master's thesis, EECS Department, University of California, Berkeley, May 2016. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/`.

Axisa, F., Dittmar, A., and Delhomme, G. Smart clothes for the monitoring in real time and conditions of physiological, emotional and sensorial reactions of human. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 4, pp. 3744–3747. IEEE, 2003.

Azar, P., Mani, A., Peng, Q., and van Gaalen, J. *Expanded Telehealth Platform for Android*. Master's thesis, EECS Department, University of California, Berkeley, May 2015. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2015/`.

Ben-Zeev, D., Schueller, S. M., Begale, M., Duffecy, J., Kane, J. M., and Mohr, D. C. Strategies for mHealth Research: Lessons from 3 Mobile Intervention Studies. *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42:pp. 157–167, 2015. ISSN 0894587X. doi:10.1007/s10488-014-0556-2.

Bloss, R. Wearable sensors bring new benefits to continuous medical monitoring, real time physical activity assessment, baby monitoring and industrial applications. *Sensor Review*, vol. 35(2):pp. 141–145, 2015.

Boulos, M. N. K., Brewer, A. C., Karimkhani, C., Buller, D. B., and Dellavalle, R. P. Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online Journal of Public Health Informatics*, vol. 5(3), 2014.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, vol. 13(3):pp. 1–17, 2011. ISSN 14388871. doi:10.2196/jmir.1838. `arXiv:1011.1669v3`.

Case, M. A., Burwick, H. A., Volpp, K. G., and Patel, M. S. Accuracy of Smartphone Applications and Wearable Devices for Tracking Physical Activity Data. *JAMA*, vol. 313(6):pp. 625–626, 2015.

Chaudhry, S. I., Barton, B., Mattera, J., Spertus, J., and Krumholz, H. M. Randomized trial of telemonitoring to improve heart failure outcomes (Tele-HF): study design. *Journal of cardiac failure*, vol. 13(9):pp. 709–714, 2007.

Chaudhry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., Phillips, C. O., Hodshon, B. V., Cooper, L. S., and Krumholz, H. M. Telemonitoring in patients with heart failure. *New England Journal of Medicine*, vol. 363(24):pp. 2301–2309, 2010.

Chen, C. and Womack, B. Google Reveals Health-Tracking Wristband at `http://www.bloomberg.com/news/articles/2015-06-23/google-developing-health-tracking-wristband-for-health-research`. June 2015. Accessed: 09/10/2016.

Chen, K. Y. and Sun, M. Improving energy expenditure estimation by using a triaxial accelerometer. *Journal of Applied Physiology*, vol. 83(6):pp. 2112–2122, 1997.

Chih, M. Y., Patton, T., McTavish, F. M., Isham, A. J., Judkins-Fisher, C. L., Atwood, A. K., and Gustafson, D. H. Predictive modeling of addiction lapses in a mobile health application. *Journal of Substance Abuse Treatment*, vol. 46(1):pp. 29–35, 2014. ISSN 07405472. doi:10.1016/j.jsat.2013.08.004. `NIHMS150003`, URL `http://dx.doi.org/10.1016/j.jsat.2013.08.004`.

Clark, R. A., Inglis, S. C., McAlister, F. A., Cleland, J. G., and Stewart, S. Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. *BMJ*, vol. 334(7600):p. 942, 2007.

Dannecker, K. L., Petro, S. A., Melanson, E. L., and Browning, R. C. Accuracy of fitbit activity monitor to predict energy expenditure with and without classification of activities. *Medicine & Science in Sports & Exercise*, vol. 43(5):p. 62, 2011.

Dhurandhar, N. V., Schoeller, D. A., Brown, A. W., Heymsfield, S. B., Thomas, D. M., Sørensen, T. I., Speakman, J. R., Jeansonne, M. M., and Allison, D. B. Energy balance measurement: when something is not better than nothing. *International Journal of Obesity*, 2014.

Donaire-Gonzalez, D., de Nazelle, A., Seto, E., Mendez, M., Nieuwenhuijsen, M. J., and Jerrett, M. Comparison of physical activity measures using mobile phone-based CalFit and actigraph. *Journal of Medical Internet Research*, vol. 15(6), 2013.

Dwork, C. Differential privacy. In *Automata, Languages and Programming*, pp. 1–12. Springer, 2006.

EMB/11073. ISO/IEEE Health informatics – Personal health device communication Part 00103: Overview. *ISO/IEEE Std 11073-00103:2012*, 2012.

Eng, D. S. and Lee, J. M. The promise and peril of mobile health applications for diabetes and endocrinology. *Pediatric diabetes*, vol. 14(4):pp. 231–238, 2013.

Galbreath, A. D., Krasuski, R. A., Smith, B., Stajduhar, K. C., Kwan, M. D., Ellis, R., and Freeman, G. L. Long-term healthcare and cost outcomes of disease management in a large, randomized, community-based population with heart failure. *Circulation*, vol. 110(23):pp. 3518–3526, 2004.

Giamouzis, G., Mastrogiannis, D., Koutrakis, K., Karayannis, G., Parisis, C., Rountas, C., Adreanides, E., Dafoulas, G. E., Stafylas, P. C., Skoularigis, J., Giacomelli, S., Olivari, Z., and Triposkiadis, F. Telemonitoring in chronic heart failure: a systematic review. *Cardiology Research and Practice*, vol. 2012, 2012.

Hamill, J., Derrick, T. R., and Holt, K. G. Shock attenuation and stride frequency during running. *Human movement science*, vol. 14(1):pp. 45–60, 1995.

Heiderscheit, B. C., Chumanov, E. S., Michalski, M. P., Wille, C. M., and Ryan, M. B. Effects of step rate manipulation on joint mechanics during running. *Medicine and science in sports and exercise*, vol. 43(2):p. 296, 2011.

Hunter, D. L. An Apple a day keeps the research ethics committee away? *Research Ethics*, vol. 11(1):pp. 2–3, 2015. ISSN 1747-0161. doi:10.1177/1747016115585299. URL `http://rea.sagepub.com/lookup/doi/10.1177/1747016115585299`.

Hussain, M., Al-Haiqi, A., Zaidan, A., Zaidan, B., Kiah, M., Anuar, N. B., and Abdulnabi, M. The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations. *Computer Methods and Programs in Biomedicine*, vol. 122(3):pp. 393–408, 2015.

Inglis, S. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. *Journal of Evidence-Based Medicine*, vol. 3(4):pp. 228–228, 2010.

Jardine, J., Fisher, J., and Carrick, B. Apple's ResearchKit: smart data collection for the smartphone era? *Journal of the Royal Society of Medicine*, vol. 108(8):pp. 294–296, 2015. ISSN 0141-0768. doi:10.1177/0141076815600673. URL `http://jrs.sagepub.com/lookup/doi/10.1177/0141076815600673`.

Klaassen, R., op den Akker, R., and op den Akker, H. Feedback Presentation for Mobile Personalised Digital Physical Activity Coaching Platforms. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, pp. 64:1–64:8. ACM, New York, NY, USA, 2013. ISBN 978-1-4503-1973-7. doi:10.1145/2504335.2504404. URL `http://doi.acm.org/10.1145/2504335.2504404`.

Lee, Y.-D. and Chung, W.-Y. Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring. *Sensors and Actuators B: Chemical*, vol. 140(2):pp. 390–395, 2009.

Martin, C. K., Miller, A. C., Thomas, D. M., Champagne, C. M., Han, H., and Church, T. Efficacy of SmartLoss(SM) , a smartphone-based weight loss intervention: Results from a randomized controlled trial. *Obesity*, vol. 23(5):pp. 935–42, 2015. ISSN 1930-739X. doi:10.1002/oby.21063. URL `http://onlinelibrary.wiley.com/doi/10.1002/oby.21063/full`.

Mladenov, M. and Mock, M. A step counter service for Java-enabled devices using a built-in accelerometer. In *Proceedings of the 1st International Workshop on Context-Aware Middleware and Services: Affiliated With the 4th International Conference on Communication System Software and Middleware (COMSWARE 2009)*, pp. 1–5. ACM, 2009.

Ong, M. K., Romano, P. S., Edgington, S., Aronow, H. U., Auerbach, A. D., Black, J. T., De Marco, T., Escarce, J. J., Evangelista, L. S., Hanna, B., et al. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the better effectiveness after transition–heart failure (BEAT-HF) randomized clinical trial. *JAMA internal medicine*, vol. 176(3):pp. 310–318, 2016.

Pande, A., Zeng, Y., Das, A. K., Mohapatra, P., Miyamoto, S., Seto, E., Henricson, E. K., and Han, J. J. Energy expenditure estimation with smartphone body sensors. In *Proc. of the 8th International Conference on Body Area Networks*, pp. 8–14. 2013.

Paré, G., Moqadem, K., Pineau, G., and St-Hilaire, C. Clinical effects of home telemonitoring in the context of diabetes, asthma, heart failure and hypertension: a systematic review. *Journal of Medical Internet Research*, vol. 12(2), 2010.

Park, J.-g., Patel, A., Curtis, D., Teller, S., and Ledlie, J. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 113–122. ACM, 2012.

Patel, M. S., Asch, D. A., and Volpp, K. G. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, vol. 313(5):pp. 459–460, 2015. doi:10.1001/jama.2014.14781. URL `http://dx.doi.org/10.1001/jama.2014.14781`.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *Biomedical Engineering, IEEE Transactions on*, vol. 58(1):pp. 7–11, 2011.

Remme, W. J. and Swedberg, K. Guidelines for the diagnosis and treatment of chronic heart failure. *European heart journal*, vol. 22(17):pp. 1527–1560, 2001.

Rickles, N. M., Svarstad, B. L., Statz-Paynter, J. L., Taylor, L. V., and Kobak, K. A. Pharmacist telemonitoring of antidepressant use: effects on pharmacist–patient collaboration. *Journal of the American Pharmacists Association*, vol. 45(3):pp. 344–353, 2005.

Ritter, S. Apple's Research Kit Development Framework for Iphone Apps Enables Innovative Approaches to Medical Research Data Collection. *Clinical Trials*, vol. 5(2):p. 1000e120, 2015. ISSN 21670870. doi:10.4172/2167-0870.1000e120.

Rowlands, A. V., Eston, R. G., and Tilzey, C. Effect of stride length manipulation on symptoms of exercise-induced muscle damage and the repeated bout effect. *Journal of sports sciences*, vol. 19(5):pp. 333–340, 2001.

Spring, B., Gotsis, M., Paiva, A., and Spruijt-Metz, D. Healthy apps: Mobile devices for continuous monitoring and intervention. *IEEE Pulse*, vol. 4(6):pp. 34–40, 2013. ISSN 21542287. doi: 10.1109/MPUL.2013.2279620.

Swedberg, K., Cleland, J., Dargie, H., Drexler, H., Follath, F., Komajda, M., Tavazzi, L., Smiseth, O. A., Gavazzi, A., Haverich, A., et al. Guidelines for the diagnosis and treatment of chronic heart failure: executive summary (update 2005). *European heart journal*, vol. 26(11):pp. 1115–1140, 2005.

Taylor, A. G. The ResearchKit Health Projects. In *Get Fit with Apple Watch*, chap. 8, pp. 111–117. Apress, 2015. ISBN 978-1-4842-1282-0.

Warburton, D. E. R., Nicol, C. W., and Bredin, S. S. D. Health benefits of physical activity: the evidence. *CMAJ : Canadian Medical Association Journal = Journal de l'Association medicale canadienne*, vol. 174(6):pp. 801–9, 2006. ISSN 1488-2329. doi:10.1503/cmaj.051351. `arXiv:1011.1669v3`, URL `http://www.ncbi.nlm.nih.gov/pubmed/16534088`.

Chapter 3

Privacy in Telemonitoring

> *"The real danger is the gradual erosion of individual liberties through automation, integration, and interconnection of many small, separate record-keeping systems, each of which alone may seem innocuous, even benevolent, and wholly justifiable."*
>
> – US Privacy Study Commission, 1977

3.1 Introduction

The study of privacy dates back to at least the ancient Greek philosophy, when Aristotle made the distinction between the *polis*–the public sphere of political activity, and the *oikos*–the private sphere of domestic life. Since the years of ancient Greece until our present day, the concept of privacy has gone through major remodeling due to changes in the ways of life. Two of the most recent updates to our collective understanding of privacy are i) the emergence of modern photography and the printed press in the 19th century; and ii) the integration of information systems into our daily lives in the 20th century. In each one of these two events, the collective view of privacy has shifted to adapt to the new technological advancements.

For instance, before modern photography, one had to be still for a significant period of time for the photographer to take one's photo. After modern photography, reporters and journalists started taking pictures of people without their consent, giving rise to then-new privacy questions. In light of the new realities imposed by modern photography and the printed press, Samuel D. Warren and Louis D. Brandeis wrote "The Right to Privacy," in 1890. In it, Warren and Brandeis attempted to define privacy, and wrote that privacy is "the right to be let alone." They further contrast how the law had broadened from the confined view of physical harm to include notions that were emotional, intellectual and of the "spiritual nature." Under that comparative narrative, they discuss the question of whether the law at the time afforded principles that can be invoked to protect the privacy of an individual. It was clear then that a new notion of invasion of privacy has emerged, and that the law had to develop in order to capture these developments.

Not 80 years later, another wave of questions regarding how the law protected our privacy emerged. In 1968, Alan Westin published his book "Privacy and Freedom," focusing on the then-new question of privacy in the age of information systems and databanks [Westin, 1968]. The wave of changes to the law as a result of introducing information systems to our lives is conceivably still happening until our present day. We will elaborate further on these changes in Section 3.3.

We are arguably living on the brink of yet another wave of brand-new privacy concerns. It is true that the information age transformed our view on privacy in a major way, but perhaps what this age is enabling us to do is an even bigger privacy challenge. In light of the ever-advancing fields of machine learning and statistical inference, more and more technologies are making use of the data we disclose in order to *infer* information about us that we did not disclose, and perhaps did not want to disclose. This, on its own right, would not constitute a major privacy concern except for the fact that these inferences are becoming accurate to a point where it is hard to ignore their influence on the privacy of the individual. To give an example of such intrusive inference, our (undisclosed) political affiliation can be inferred from our (disclosed) TV show ratings and viewing habits [Salamatian et al., 2013]. Some research that attempts to reduce the risk of information leakage through inference attacks has been done but we are far from home [see Reed, 1973; Yamamoto, 1983; Evfimievski et al., 2003; Dwork, 2006; Rebollo-Monedero et al., 2010; du Pin Calmon and Fawaz, 2012; Sankar et al., 2013, for such examples].

In our view, one reason privacy-aware technologies are not gaining much positive momentum among top companies in the industry is the inherent trade-off between privacy and utility. It is often the case that increasing privacy protection–in statistical inference settings–translates to lower utility of data. This result has even been formalized in the context of statistical databases to argue that full privacy can only be achieved when no utility is obtained from the database [Dwork, 2006]. To explain this result intuitively, if a statistical database wants to preserve full privacy, it can answer every query completely randomly, independent of the data stored inside it (e.g., flip a coin and answer 0 if heads and 1 if tails, regardless of the query or the content of the database). Alternatively, the database can answer deterministically, also independent of the data stored within (e.g., output -3 to every query regardless of the query or the content of the database). Both of these strategies obtain full privacy because they reveal absolutely nothing about the data stored inside the database. However, in the same breath, these strategies eliminate any utility the database may have had. It is then desirable to design privacy frameworks and mechanisms, that offer a good privacy-utility tradeoff.

Since inference algorithms are the backbone of the predictive healthcare model, it is important to systematically understand this privacy threat and attempt to mitigate it. In this chapter, we address this very question. The rest of this chapter is organized as follows. In Section 3.2, we introduce preliminary evidence that consumers trust technology researchers with their health data to comparable level of trust that they hold towards their physicians. We argue that this positions technology researchers in a unique spot of power and responsibility that they need to harness in order to build privacy-aware predictive medicine systems. Then, we survey the existing privacy design principles and practices in Section 3.3 and argue for the need of an additional principle. Afterwards, we introduce the framework for Private Disclosure of Information (PDI), which aims at preserving the privacy of individuals in the process of disclosing their information against Man-In-The-Middle (MITM) inference attacks. We show that, in our setting, full privacy *can* be achieved while maintaining full utility of the data to its intended recipient. Although PDI is a general framework that may be applied in applications other than predictive medicine, we describe it in a language consistent with predictive medicine, the theme of this dissertation. Finally, we discuss our findings and summarize this chapter in Section 3.5.

The contributions of this chapter are i) preliminary evidence to the acceptability of telemonitoring technologies in different applications (Section 3.2); ii) introducing the *Inference* privacy design principle (Section 3.3.2); and iii) introducing the framework for PDI, alongside a MATLAB toolbox that implements its learning mechanism (Section 3.4).

3.2 User Privacy and Acceptability of Telemonitoring

3.2.1 Introduction

In each of the protocols of the studies that we conducted for the Berkeley Telemonitoring framework–discussed in Chapter 2–we administered a privacy and acceptability survey. In the survey, we attempt to understand i) the level of acceptability of these technologies to the consumer; ii) the perceived efficacy of these technologies by the consumer; iii) the consumers' privacy-related concerns regarding these technologies; and iv) the privacy requirements of these technologies. One of the dimensions of interest concerns the trust level of the target population in technology researchers that are studying telemonitoring technologies.

This marker is important because it allows us to assess the feasibility of designing telemonitoring

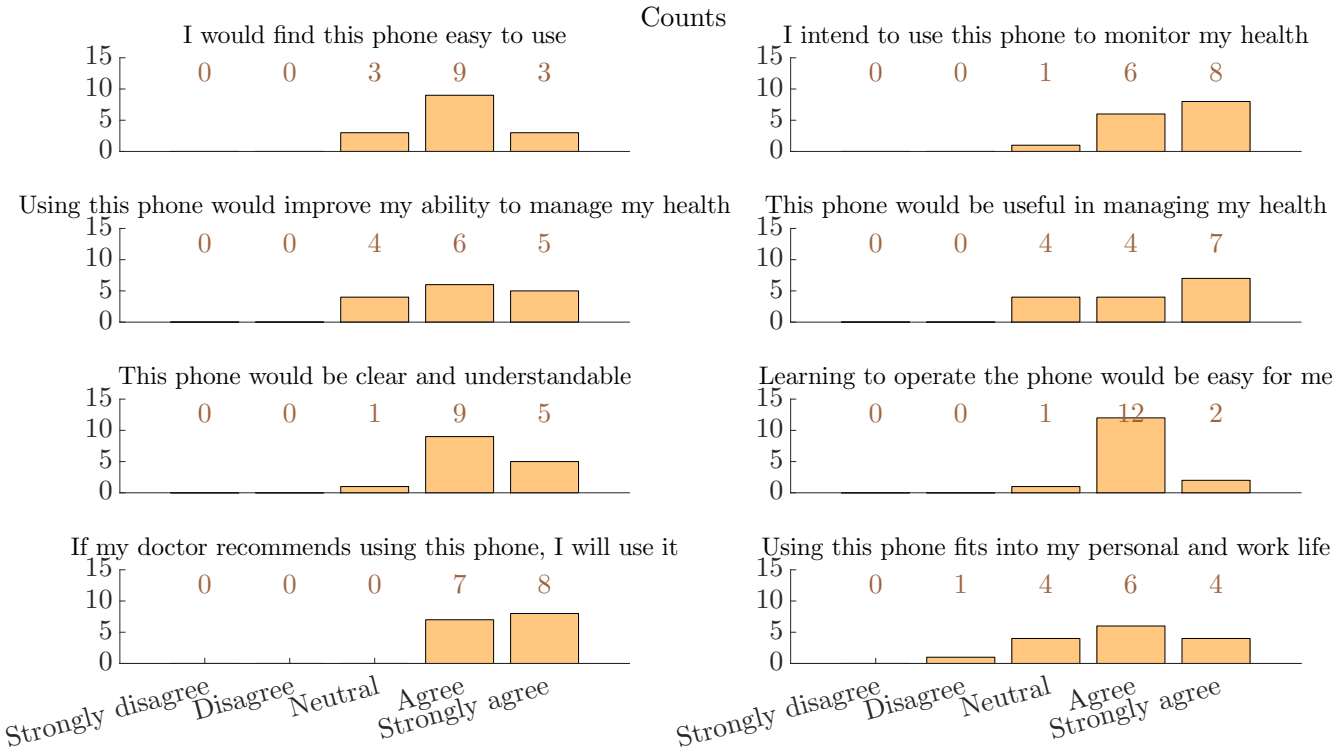


Figure 3.1: Pre-study acceptability survey results for the CHF study (N=15).

systems with high efficacy. Understanding this trust level is also important because it embodies within it a level of responsibility that we, as technology researchers, carry on our shoulders. It is our responsibility to design privacy-aware systems and technologies, and whatever level of trust we may receiver, we shall not lose. With this background, we present excerpts from the results of the privacy and acceptability surveys from two of our studies.

3.2.2 CHF – Privacy and Acceptability Study

In this section, we discuss the acceptability and privacy findings of the congestive heart failure (CHF) study that we discussed in detail in Section 2.4 [Aranki et al., 2016]. Although this part of the dissertation is self-contained, we refer the reader to Section 2.4 for more context and details about the study. The study remotely monitored 15 patients with CHF for a period of 3 months. The study was conducted in collaboration with the Northwestern Medical Faculty Foundation in Chicago, IL. All 15 subjects were patients of Northwestern Memorial Hospital in Chicago, IL.

As part of the study, we wanted to assess the acceptability of telemonitoring for CHF purposes. In order to achieve this assessment, the protocol included a survey that would be taken by the participating subjects before (pre) and after (post) the study. The survey instrument that was used in this study was designed in collaboration with Heather M. Patterson, Martin French and Helen Nissenbaum. The framework for contextual integrity was utilized in the design of the instrument [Nissenbaum, 2009].

All 15 participating subjects took part in the pre-study acceptability survey. The responses of the acceptability portion of the pre-study survey are summarized in Figure 3.1. In the figure, we present the number of subjects who selected each one of the five possible answers to each of the acceptability questions. From the results, there is evidence that the majority of the participating

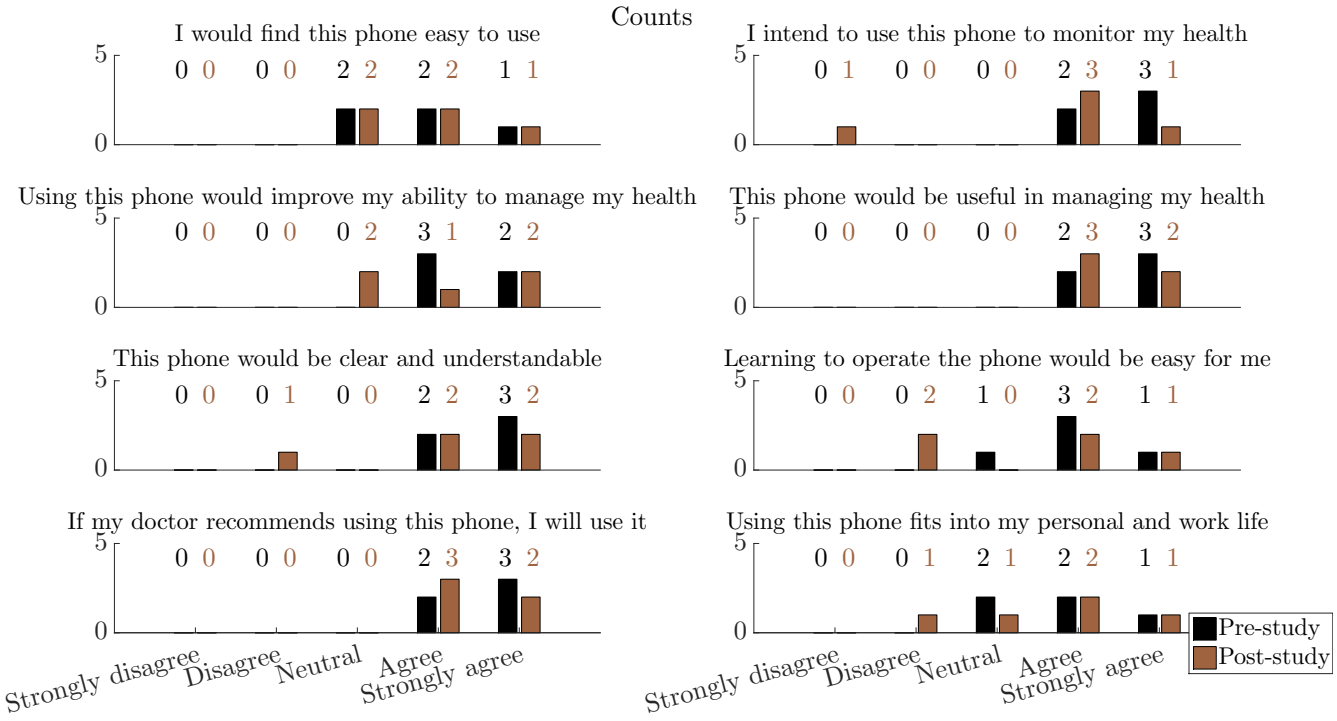


Figure 3.2: The acceptability survey results, in pre- and post-study surveys, for the 5 subjects who completed the CHF post-study survey.

subjects were receptive and accepting of telemonitoring for CHF purposes prior to the study.

The post-study survey was administered with the same questions as in the pre-study version. Only 5 subjects took part in the post-study survey. In Figure 3.2 we present the responses of these 5 subjects to the same acceptability questions, in both pre- and post-study surveys. Similar to before, we present the number of subjects who selected each of the five possible responses to each of the questions. We note that even though the number of subjects that participated in both the pre- and post-study surveys is too small to make any general deductions, there is a minor trend of lower acceptability after the study. Such a trend is expected since the study was exploratory and was less focused on user experience. It is also worth noting that most changes in the acceptability answers are small (one level up or down) and that the post-study results still indicate positive acceptability levels.

As for the privacy side of the survey, the following statement was presented to the subjects: "*Sometimes the cell phone might automatically record, or ask you to report, specific kinds of information about your health or behavior, such as your weight, your mood, or your blood pressure. The following questions will help us understand how comfortable you are with the idea of other people knowing these things about you.*" The subjects were specifically asked to indicate their comfort level in sharing their weight, level of physical activity, exact location and blood pressure with i) their doctors or nurses who treat their heart failure; ii) researchers who study health care telemonitoring technologies; iii) public health professionals who study the causes of heart diseases; iv) insurance companies that set their health insurance policies; and v) their close family members who take care of them at home [Nissenbaum, 2009; Aranki et al., 2016]. We present the 15 subjects' responses to this question, in the pre-study survey, in Figure 3.3.

Perhaps unsurprisingly, from Figure 3.3 we observe that the subjects' level of comfort in sharing their data with their doctors or nurses is high. Perhaps more interestingly, the subjects' responses

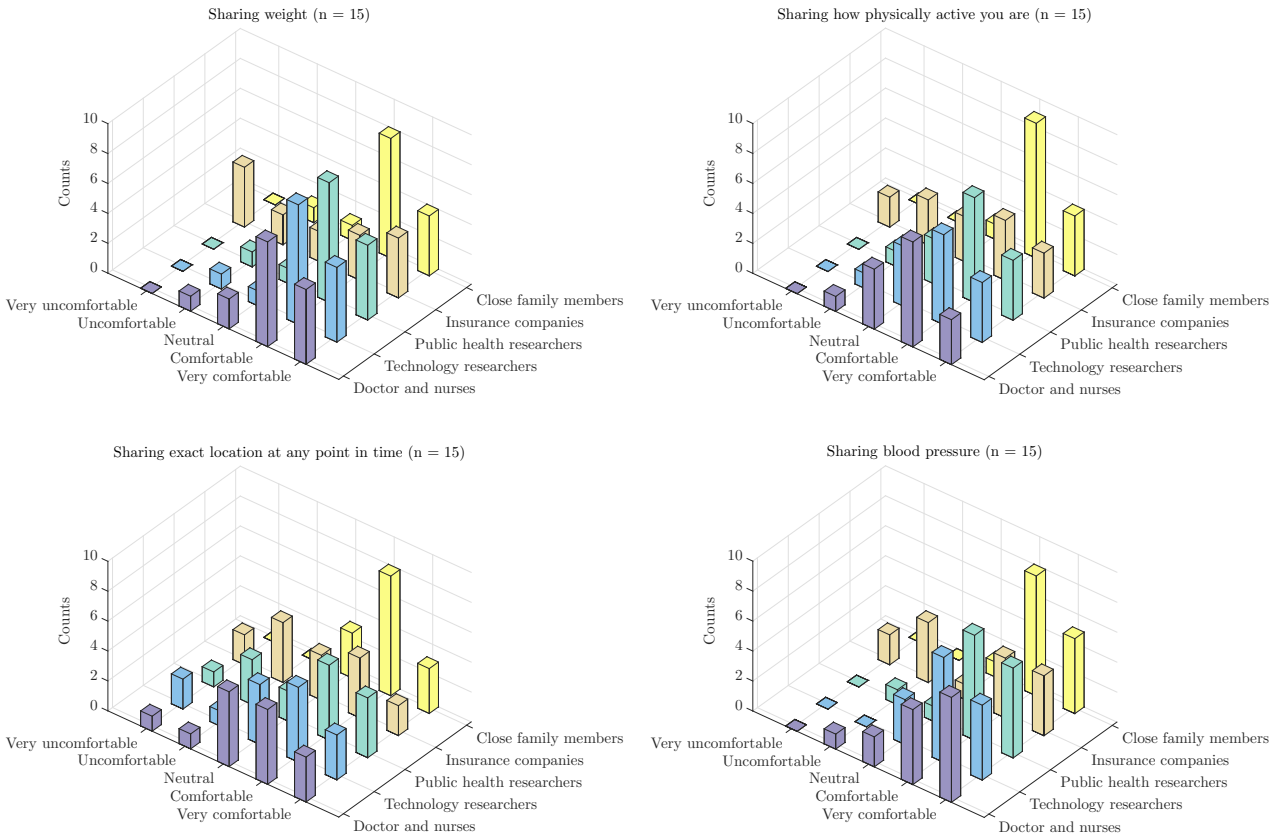


Figure 3.3: The subjects' responses to the pre-CHF-study survey privacy-related questions administered, inquiring about how comfortable they are when sharing their data on weight, physical activity levels, location, and blood pressure ($N = 15$).

regarding their level of comfort in sharing data with technology researchers is comparable to those of sharing their data with doctors and nurses (and to those of sharing data with public health researchers). The subjects' comfort levels in sharing data with technology researchers is also comparable to this of close family members in all cases except for sharing exact location. Finally, we observe that the subjects' level of comfort in sharing data with technology researchers is higher than the level of comfort in sharing the same data with insurance companies. In conclusion, the results in Figure 3.3 indicate that the subjects are comfortable using the proposed telemonitoring technology from a privacy point of view.

Moreover, these results load some responsibility on our shoulders as technology researchers. First, there is the inherent responsibility of not losing this trust. Perhaps not any less importantly, with this level of trust, not only do we have a main role in shaping the technology in a privacy-aware manner, but we also have the means to achieve it by having direct access to consumers who trust us to do so.

Next, we provide similar evidence of consumer trust from a later but smaller study that we conducted, called RunningCoach.

3.2.3 RunningCoach – Privacy and Acceptability Study

In this section, we discuss the acceptability and privacy findings of the RunningCoach study that we discussed in detail in Section 2.8 [Aranki et al., 2017]. This part of the dissertation is also

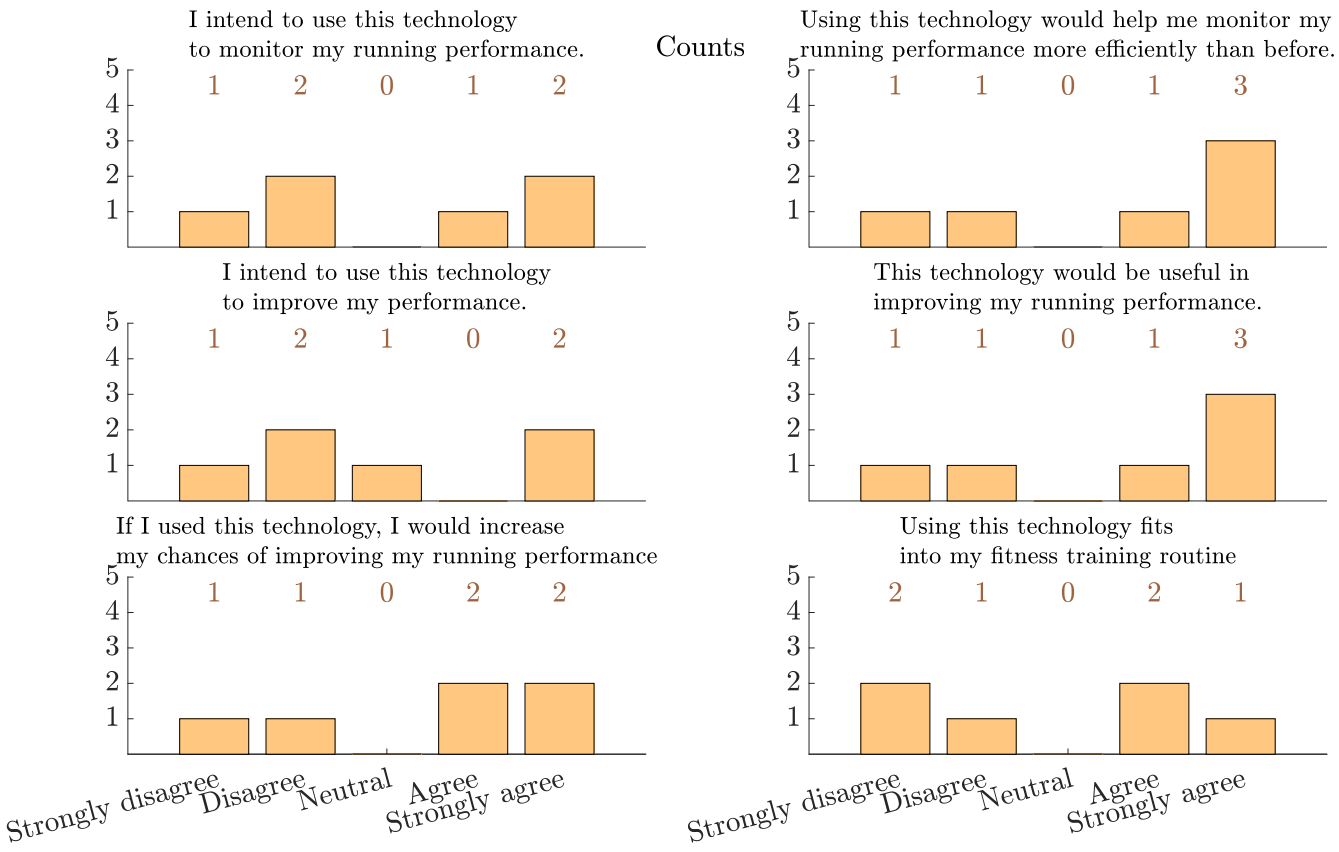


Figure 3.4: Post-study acceptability survey results for the RunningCoach study (N=6).

self contained; nevertheless, we refer the reader to Section 2.8 for more context and details about the study. The study remotely monitored 6 long-distance runners for a period of 3 months. The study was conducted as part of the Berkeley Telemonitoring project's efforts to demonstrate the proof-of-concept of the Berkeley Telemonitoring framework. All 6 subjects were students at UC Berkeley. The survey instrument that was used in this study was inspired by and adapted from the survey instrument in the CHF study.

In this study, the subjects only took the survey after they finished the study. In Figure 3.4 we present the subjects' responses to the acceptability portion of the survey. We observe that the acceptability of telemonitoring for long-distance running coaching is lower than the acceptability of telemonitoring for CHF. There are multiple factors that could have played a role in this difference. First, from interviews with the subjects, it was evident that they prefer a smaller form factor monitoring device than a smartphone. One subject stated: "I'm a big fan of using running watches instead of phone apps because the form factor is much more comfortable. That's the main reason I was so negative about using a phone-based athletic trainer." It is important to mention here that we are able to avert the challenge of a large form factor because the Berkeley Telemonitoring framework supports Android devices in general. As such, a telemonitoring application for, say an Android smart watch, can be implemented using the Berkeley Telemonitoring framework. Second, it is not a stretch to hypothesize that the higher the perceived utility, the higher the price people are willing to pay for it. From that point of view, CHF telemonitoring application has a potentially higher perceived utility and value, which encourages people to be more willing to adopt it.

The highlight of the survey results, though, lie in the privacy part of it. Similar to the CHF study, the survey asked subjects the following question. "*Sometimes the smartphone might auto-*

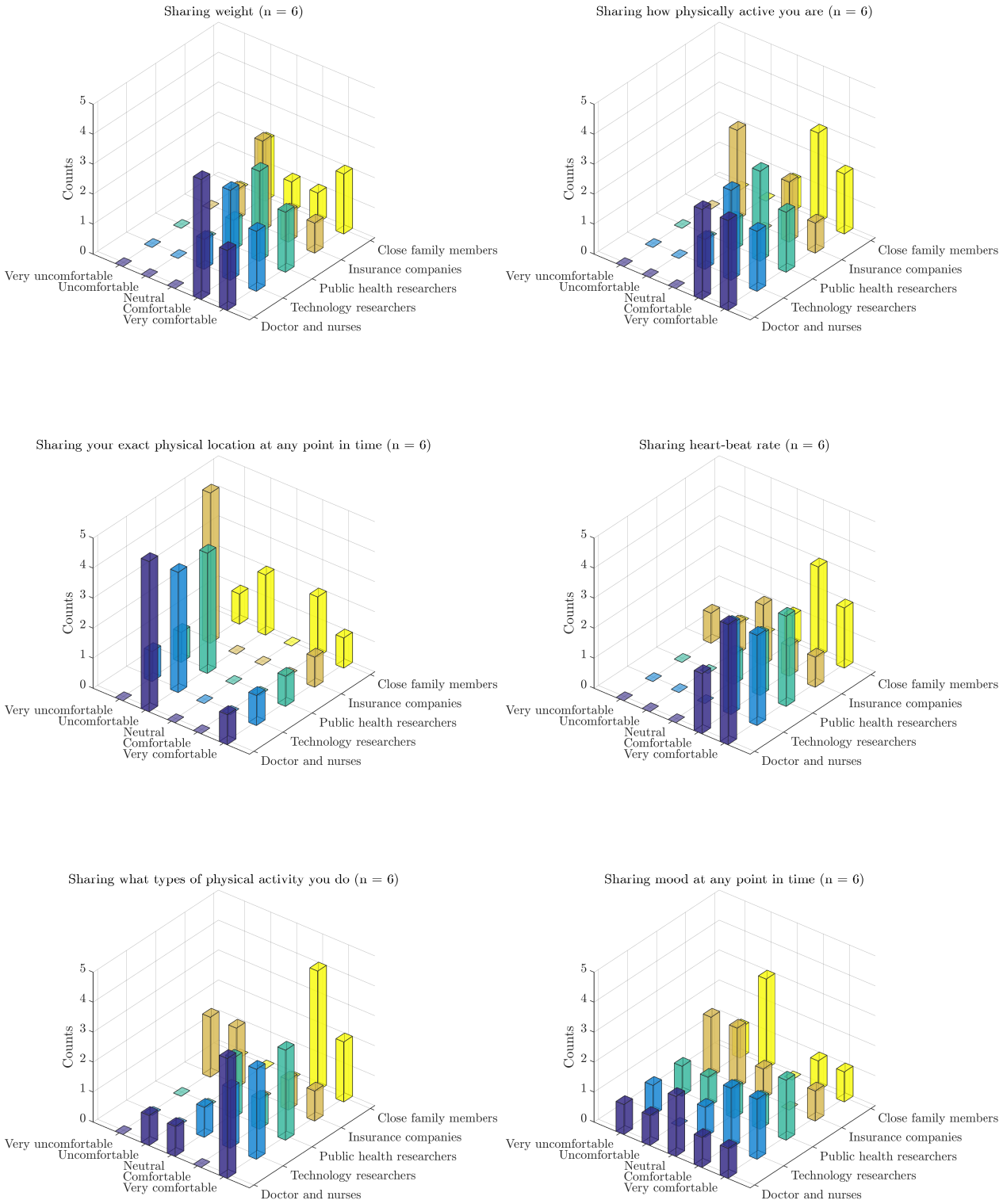


Figure 3.5: The responses to privacy related question administered after to the RunningCoach study, inquiring about how comfortable the users are sharing their data on weight, physical activity levels, location, and heart-beat rate values (N=6).

matically record, or ask you to report, specific kinds of information about your health or behavior, such as your weight, your mood, or your blood pressure. The following questions will help us understand how comfortable you are with the idea of other people knowing these things about you."

We show here the subject's level of comfort in giving regular updates about their i) weight; ii) level of physical activity; iii) exact physical location at any point in time; iv) heart-beat rate; v) types of physical activity she or he does; and vi) mood at any point in time to i) doctors and nurses who provide her or his healthcare; ii) researchers who study athletic training technology; iii) public health professionals who study the effects of exercise and athleticism; iv) insurance companies that set her or his health insurance prices; and v) close family members who care about her or his health.

In particular, Figure 3.5 shows findings that are consistent with our preliminary evidence from the CHF study. The data suggest that i) the subjects here also trust technology researchers with their health, fitness, GPS and mood data at a comparable level to their trust in their physician; and ii) there is a general privacy acceptability of telemonitoring for fitness and coaching purposes.

3.2.4 What's Next?

First, we showed that the telemonitoring technologies are acceptable from a privacy point of view in two applications. Moreover, we showed data from our studies that suggest consumers' have trust in technology researchers when it comes to sensitive data. This preliminary evidence was corroborated in a second study with a different application. But what does this mean? We argue that this puts a heightened responsibility on the shoulders of technology researchers but also gives them a unique opportunity to develop privacy-aware technologies. So what can we do? We will first survey the popular principles in design of privacy-aware systems and augment these principles to include newer threats that need to be addressed. Afterwards, we will introduce the framework for Private Disclosure of Information that aims at defending the subject from a MITM passive inference attack.

3.3 The Inference Design Principle

We have established that privacy is an important factor in designing any system, in particular those applicable to predictive medicine. We have also outlined preliminary evidence that in a range of applications of telemonitoring, the level of trust in technology researchers is comparable to that in medical personnel. This puts technology researchers and designers at a unique place, given the access they have to such sensitive data. We turn to discuss two important questions. Namely, i) what design principles should we adopt in designing telemonitoring applications? And ii) what are the privacy threats in telemonitoring settings?

We will answer the first question by identifying a new design principle for privacy-aware systems that addresses the new privacy threats of statistical inference. We introduce this *Inference* principle in Section 3.3.2. But before we do, we need to provide context by first outlining the existing principles and practices in privacy-aware system design, which we do in Section 3.3.1.

3.3.1 Existing Privacy Design Principles and Practices

What principles should technology researchers and designers adopt when designing and implementing predictive medicine technologies in general, and telemonitoring systems in particular? In order

to answer this question, we must first review and understand some of the established principles in privacy protection as laid out by privacy scholars and adopted by different regulatory agencies. A whole book can be written to review such efforts; therefore, for the sake of brevity and conciseness, we will briefly review 4 of the more prominent and relevant efforts in that realm. These are, i) the works of Westin [1968]; ii) the United States of America (US) Privacy Act of 1974; iii) the European Union (EU)'s Directive 95/46/EC of 1995; and iv) the Federal Trade Commission (FTC) Fair Information Practice Principles of 1998.

Afterwards, we will describe 2 examples of distilled engineering practices, based on the aforementioned principles. These efforts are i) the World Wide Web Consortium (W3C) Privacy Preferences Platform (P3P) specification, which allows Web developers to write their privacy policies in a machine readable way, using Extensible Markup Language (XML) tags; and ii) Privacy by Design, which outlines a set of practices for designing privacy-aware ubiquitous systems. In what follows, we will elaborate briefly on these principles and practices.

Alan Westin's Framework for Privacy in Personal Information

Ever since information technology systems started entrenching in our daily lives, the collection, analysis, retention and dissemination of personal information has become ubiquitous at an ever decreasing cost. This led scholars and legislative bodies to examine the desired principles that are to be followed in the then-new realm of unprecedented data collection in ever growing quantity and detail. One of the pioneers of data privacy research was Alan Westin, whose study of data privacy prompted US privacy legislation [Privacy Act, 1974] as well as other privacy-advocacy movements around the world [Westin, 1968]. Therefore, no modern privacy textbook relevant to information technology can be complete without visiting, even if briefly, the distilled framework that is based on Westin's work.

Westin's contributions include 30 privacy-related surveys that were conducted between the years 1974 and 2004 [Westin and The Staff of The Center for Social & Legal Research, 2003]. Among these, Westin surveyed public opinion on health-related privacy, credit reporting and privacy, and e-commerce related privacy.

His scholarly work further resulted in shaping a framework of principles for the collection, use, retention and dissemination of personal information. These principles are often called "fair information practices." They were later adopted, in an adapted form, by the FTC in the Fair Information Practice Principles which will be discussed further in this section [Pitofsky et al., 1998, 2000]. Before we get to that, let us first examine an earlier effect of Westin's work, the US Privacy Act of 1974 [Privacy Act, 1974].

The US Privacy Act of 1974

There are two historical contexts that are relevant to the enactment of the US Privacy Act of 1974. First, in light of the Watergate scandal, the US Congress was eager to restrain the illegal investigation and surveillance of individuals by federal agencies [Department of Justice's Office of Privacy and Civil Liberties (OPCL), 2015]. Second, there was a general concern of potential abuse given the increasing use of computer systems and databanks to retain and disseminate information about individuals. As a result, in an effort to regulate the collection, use, retention and dissemination of personal information, the US enacted the Privacy Act in 1974 that sets forth principles for federal government agencies regarding their collection of information about individuals.

It is important to note that although the US Privacy Act of 1974 protects every individual subject to the collection of personal information, it only applies to records held by government agencies. We first highlight that an *individual* in this context refers only to persons who are US citizens or "aliens lawfully admitted for permanent residence" [Privacy Act, 1974]. Second, since the US Privacy Act of 1974 only applies to federal agencies, it lacks coverage in the commercial arena. As a result, the US Privacy Act of 1974 does not offer protection for consumers from the reach of corporation data collection that is very prominent in our days. The FTC has attempted to fill this gap in their Fair Information Practice Principles that will be discussed further in this section.

In spite of its failure at home, the US Privacy Act of 1974 scored some important feats. First, in its statement of purpose, the US Privacy Act of 1974 declares that "the right to privacy is a personal and fundamental right protected by the Constitution of the United States" [Privacy Act, 1974]. Not only is it a major step to explicitly identify privacy as a right protected by the US Constitution, but this statement also extends the scope of privacy protection to the age of information systems.[1]

In order to understand the second important contribution of the US Privacy Act of 1974, we need to describe its principles. These principles, called the *fair information practices*, are based on the work of Westin. They can be summarized as follows [Langheinrich, 2001].

1. Openness and transparency: neither the collection of personal information, nor the types and nature of data collected, may be done in secret.

2. Individual participation: any person who is the subject of data collection should be able to view the records pertaining to her or him, and be able to correct such records.

3. Collection limitation: the extent of the collection of personal information should be proportional to the purpose of such collection.

4. Data quality: the collected personal information should remain up to date for the purposes of the collection.

5. Use limitation: access to the collected personal information should be limited to authorized personnel; and the use of the collected personal information should be limited to the specific purpose of its collection.

6. Reasonable security: records of personal information should be safeguarded by proper and adequate security measures.

7. Accountability: the collectors and curators of personal information are accountable for compliance with the law and other principles.

Even though the coverage of these principles was limited in the US, the fair information practices affected every other major privacy legislation in the democratic world [Langheinrich, 2001]. Sector-specific laws were subsequently passed in the US for very limited and specific needs. Examples include the US Video Privacy Protection Act of 1988, the US Computer Matching and Privacy Protection Act of 1988 and the US Family Education Rights and Privacy Act of 1994. However, it wasn't until 1995 that another world-wide influential legislative action was taken. This is EU Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data ("Data Protection Directive").

[1]The first information technology data protection act, Bundesdatenschutzgesetz (BDSG), was actually first enacted in the German state of Hessen in 1970.

The EU Directive 95/46/EC of 1995

The EU Data Protection Directive of 1995 contained in essence a refined version of the principles of fair information practices that the US Privacy Act of 1974 introduced [European Parliament, 1995]. However, article 7(a) of the EU Data Protection Directive of 1995 dictates that personal data may be processed only if "the data subject has unambiguously given his consent." In this context, *consent* means "any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed." In contrast, the notion of consent in the US Privacy Act of 1974 is a weaker requirement because i) it is subject to more exceptions than the EU Data Protection Directive of 1995 permits; and ii) it is not as explicit as its counterpart in the EU Data Protection Directive of 1995 (e.g., does not require that consent be informed). In essence, this requirement in the EU Data Protection Directive of 1995 prevents all types of personal data collection, unless the subject has provided unambiguous informed consent to the processing of data; except for contractual and legal purposes as outlined in article 7 [European Parliament, 1995].

In addition to the the notion of unambiguous and informed consent, it is important to note that the scope of the EU Data Protection Directive of 1995 shields all natural persons, including non-EU citizens or legal residents, from the reach of any data collector, processor, recipient or third party, including a "natural or legal person, public authority, agency, or any other body" [European Parliament, 1995]. Even more, article 25(1) of the EU Data Protection Directive of 1995 dictates that personal data collected within the EU can be shared with non-EU third countries only if "the third country in question ensures an adequate level of protection." In contrast with the US Privacy Act of 1974, the EU Data Protection Directive of 1995 is far more protective and inclusive.

The EU Data Protection Directive of 1995 proved successful as it prompted many countries around the world to update their privacy legislation in order to comply with its provisions; an effect that the US failed to achieve. Subsequently, there has been some effort in the US to include broad consumer-level privacy protection practices. This effort, undertaken primarily by the FTC, is the subject of our next discussion.

The FTC Fair Information Practice Principles of 1998

In an effort to keep up with the technological advancement and the updated means of conducting business through virtual mediums, the FTC started studying online privacy in 1995. Compared to the EU, the US legislative attitude towards online consumer privacy protection has largely been i) against imposing a blanket regulation on the industry; and ii) in favor of a more industry-driven self-regulatory approach. This attitude was also originally supported by the FTC in their report to US Congress on Online Privacy in 1998. The FTC's report in 1998 found that the federal government has limited authority over online personal data collection and dissemination. Furthermore, in the report, the FTC encouraged the use of self-regulation in the industry as means to address consumer concerns regarding their online privacy; even though it had reported that no such "effective self-regulatory system" was emerging in the industry, at that time. The FTC did recommend that US Congress use legislation the area of children's online privacy [Pitofsky et al., 1998].

However, in the FTC report to US Congress on online privacy of 2000, there was a major twist and shift in the FTC's view and recommendation, in light of the lack of progress the industry had taken in self-regulating personal data collection. The report recommended legislative action, citing the "industry's limited success in implementing fair information practices online, as well

as [the] ongoing consumer concerns about Internet privacy." The report was issued with a 3-2 Commission vote, with one commissioner (Swindle) strongly dissenting and another concurring in part and dissenting in part. Commissioner Swindle's dissent indicates his disagreement with the recommendation for legislative action, citing "an unwarranted reversal of [the Commission's] earlier acceptance of a self-regulatory approach." [Pitofsky et al., 2000] This split somewhat signifies the Byzantine nature of the debate on regulating consumer online privacy in the US. Although the FTC has had successes in the past bringing legal action in cases relating to online privacy, it has primarily done so relying on the FTC Act of 1914 and its prohibition of unfair and deceptive practices [US Congress, 1914]. It seems unfair that the FTC is only given old weapons to fight modern battles: in a way this resembles "bringing a knife to a gun fight."

Despite all of that, the FTC had presented, since its 1998 report, the set of fair information practice principles of i) notice/awareness; ii) choice/consent; iii) access/participation; iv) integrity/security; and v) enforcement/redress. These principles are adapted from the framework of the US Privacy Act of 1974, which are in turn based on the work of Westin. The difference is that these principles are not limited to government agencies as data collectors, as is the case in the US Privacy Act of 1974. The fair information practice principles can be defined as follows [Pitofsky et al., 1998].

Notice/Awareness Consumers need to be notified and aware of any collection of their personal information online. Furthermore, consumers need to be aware of the collector's information practices, including but not limited to i) who is collecting the data; ii) how are the data being used; iii) who is receiving the data; iv) the means of the data collection; v) whether data collection is voluntary or required; and vi) the measures taken to ensure confidentiality, integrity and quality of the data.

Choice/Consent These principles revolve around giving the consumers choice regarding the sharing of their personal information, and the uses thereof. Particularly, consumers should have the choice to allow or deny secondary uses of the information that surpass the declared immediate purpose of the collection. Typically, these choice options are 'opt-in' and 'opt-out'. In an 'opt-in' worldview, consumers need to affirmatively accept the secondary uses of their personal information (i.e., the secondary use is not permitted by default); whereas 'opt-out' requires that consumers affirmatively decline the secondary uses of their personal information (i.e., the secondary use is permitted by default). It is worth noting that even though these principles were introduced by the FTC after the EU Data Protection Directive of 1995, the FTC consent principle is relatively less powerful than its counterpart from the EU Data Protection Directive of 1995. For example, according to the FTC Fair Information Practice Principles, consumers may have an unfair say in the consent process. Consider the following scenario for illustration. Consumers often consent to having their personal information shared with third-parties under certain circumstances. It is rare that these certain conditions are explicitly specified in the agreement. Moreover, once the third-parties gain access to such personal information, they may share it with their subsidiaries. This fundamentally nullifies the ability of the consumer to control access to their data in an informed manner [Tavani and Bottis, 2010].

Access/Participation The consumers should have access to not only view the collected data pertaining to them, but to also verify and contest its accuracy and/or completeness. The principle also dictates that for access to be meaningful, it must be available in a timely and inexpensive

manner. In addition, contesting and correcting inaccurate and/or incomplete information must be allowed through a simple mechanism that also allows the data collector to validate it. Finally, any such edits to personal information due to a subject contesting the accuracy and/or completeness of her or his data has to be then communicated to all recipients of such data.

Integrity/Security The integrity of the data means that the data be accurate for the purposes of its collection. The FTC Fair Information Practice Principles dictate that information collectors must take reasonable steps in order to ensure the integrity of the data they collect, including but not limited to: using reputable sources of data, providing consumers access to their data, and removing out-of-date data or anonymizing them.

Enforcement/Redress In order to ensure the effectiveness of any set of principles, including the aforementioned Fair Information Practice Principles, there needs to be a mechanism to enforce them. In addition, redress protocols and policies also need to be present in order to correct violations. The FTC report of 1998 enumerates 3 possible mechanisms of enforcement and redress: i) self-regulation; ii) private remedies; and iii) government enforcement.

We argue that enforcement is the pillar upon which the effectiveness of rest of these principles rests. For example, Langheinrich [2001] writes that "if certain legal requirements are simply not enforceable, technological or procedural solutions need to be found, or the law changed." The views taken in this dissertation will be discussed in more detail in Section 3.3.2. However, we observe that like is so often the case, the answer to the enforcement mechanism probably lies between the different options. However, in retrospect it seems that the approach taken by EU Data Protection Directive of 1995 is closer to the answer than the approach currently taken in the US regarding consumer privacy protection and online privacy protection.

In light of the different regulatory frameworks for consumer privacy, we now turn to survey the principles that drive the design of privacy-sensitive systems. For this purpose, we first describe two such frameworks, and then augment them with a new principle that is applicable to the new threats of statistical inference.

The W3C Privacy Preferences Platform

Given the slow adaptation of many regulatory bodies, the onus of making progress of any tangible effect now lies on the shoulders of technology developers and the privacy advocates in the industry. One trial of such effort is the World Wide Web Consortium (W3C) Privacy Preferences Platform (P3P), which is a specification that enables websites to describe their privacy policies in a machine readable manner. This enables Web browsers to parse these policies. In this scenario, Web browsers can allow users to set what policies are acceptable to them and which are not. The Web browsers can check visited website policies against these preferences, and alert the user if the website has privacy policies that are not acceptable to the user. For example, the user can request to reject any privacy policy that shares her or his shopping preferences with third parties.

Designing the P3P specification was carried by a W3C working group that included representation from privacy advocates, industry partners and universities [Cranor et al., 2006]. The P3P specification defines the syntax and semantics of a set of XML tags and elements that can be used to define policies regarding i) subjects' *access* to their data; ii) resolving *disputes* regarding the collected data (e.g., court, independent organization, etc.); iii) the *purposes* of data collection (e.g., telemarketing, contact, research and development, etc.); iv) the *recipients* of the collected data; v) the *retention* of the collected data (e.g., indefinitely, as long as needed for legal purposes,

etc.); and vi) the types and *categories* of collected data (e.g., financial, purchase, health, etc.) and whether each collected data variable is required or optional.

This effort did not achieve its full potential due to two enforcement-related issues. The first enforcement issue lies in incentivizing websites (or forcing them, through legislation) to use this mechanism to communicate their policies. The second complication of enforcement is ensuring the accuracy of these policies with respect to the actual handling and use of the data.

Let us examine a more promising privacy-aware design framework, called Privacy by Design.

Privacy by Design

The notion of Privacy by Design [Langheinrich, 2001; Schaar, 2010] and the related notion of privacy-enhancing technologies [van Rossum, 1995] came to life in response to rapid deployment of seemingly innocuous information technologies in our lives. Such systems range in their form factors, from large databases to small and almost invisible systems (e.g., ubiquitous systems, Internet of Things (IoT)). Both of these efforts attempt to set-forth a framework for the design of information systems in a privacy-aware manner. In this discussion we will focus on the framework for Privacy by Design since it was devised and refined in the field of ubiquitous computing, which is relevant for predictive medicine systems in general and telemonitoring systems in particular. It is important to note that Privacy by Design is still a design framework that is applicable to systems other than in ubiquitous computing and IoT.

In order to better describe Privacy by Design, we need to provide the context of ubiquitous computing. Ubiquitous computing is a branch of computer sciences where computing is designed to be seamless. The goal of this paradigm of computing is generally sensing for the purposes of automating tasks in our daily lives (as well as in industrial applications). As a byproduct of this sensing, ubiquitous computing allows *memory amplification* by essentially recording more and more of our environments and our lives in a way that may leave a permanent record of our physical and mental worlds. Langheinrich [2001] summarizes the 4 properties of ubiquitous computing, that make it a prime area for the study of designing privacy-aware system:

Ubiquity Ubiquity is the explicit goal of this branch of computing, being everywhere. From even the first principle of ubiquitous computing, it is clear that any decision made during the design of such artifacts will have a broad affect on our privacy. The reach of this effect includes i) our lives in the public sphere–such as driving or walking in public; ii) our lives in shared, semi-public spheres–such as the workplace or school; and iii) our lives in the private sphere–such as our homes or even ultimately, our emotional states.

Sensing One of the primary tasks of a ubiquitous system or artifact is its ability to sense some aspects of the environment accurately. For example, think about an artifact that is able to record humidity, temperature, carbon monoxide and carbon dioxide levels. As technology and research progresses, so does our ability to i) to increase the accuracy of the sensing abilities of ubiquitous systems; and ii) to sense new variables in the environment (e.g., mood, stress, emotions). If anything, this principle is the atomic reason for the privacy concerns in this field, for this principle dictates that ubiquitous computing needs to be able to collect data about the environment, including us. Even if the collected data are not, in their own right, privacy sensitive, *inferences* based on them may be.

Invisibility In addition to being everywhere, ubiquitous artifacts should become invisible i) by becoming smaller and smaller (in form factor); and ii) by requiring less and less active input from the people in the environment (ultimately none). Once again, this principle magnifies even more the privacy concerns and the importance of being sensitive to these issues in the design of such systems. This is primarily because it is easy to imagine a scenario where ubiquitous systems are so invisible that individuals don't even know about their existence; much like how data collection on the Web is hard to quantify accurately because it is invisible to us to a large extent.

Memory amplification Memory and storage are becoming cheaper and cheaper. Since ubiquitous systems should have sensing abilities and are designed to be everywhere, they can eventually result in prolonging the lifetime of events in our lives. For instance, it is not a science fiction exercise to picture the world where it is impossible for someone to do something without being recorded, with this record potentially being retained for a time longer than one's life expectancy. This amplification of the collective memory in turn amplifies the privacy concerns in ubiquitous computing even more.

It is important to highlight that telemonitoring is an example of a ubiquitou system, and therefore potentially suffer from all of the aforementioned privacy concerns, and more. Moreover, it is vital to understand that decisions made in the design process will have their effects on viable policies that can be legislated. The primary reason for this entanglement is that these design decisions dictate what later can be enforced (e.g., by law). For example, passive Global Positioning System (GPS) tracking devices rely on estimating their distances from known satellites. These devices achieve that by passively listening to signals coming from these satellites. This design implies that a company–say, the deploys satellites–cannot later charge you a regular service fee for using their satellites, because they cannot know that you are using their satellites. In this example, it is clear that the specific design decisions made during the design of GPS affected what policies can later be enforced.

If we combine the importance of enforcement with the fact that design decisions and choices affect enforceability, then we arrive at the conclusion that technology designers have to consider privacy and enforceability when they design information systems in general, and ubiquitous systems in particular (including those applicable for predictive medicine such as telemonitoring). The previous statement is the essence of the framework for Privacy by Design and its vitality.

With this context in mind, we enumerate the 6 principles of Privacy by Design as described by Langheinrich [2001].

Notice/Openness According to Langheinrich, the most fundamental principle of any data collection system is openness (or notice). This principle dictates that a data collection system should notify the subjects of such collection that their information is being collected. We note that P3P is a system that is designed, in part, to enable websites to achieve this principle in a fair manner. Ideally, such an effort needs to happen during the design of the system, not as a post-hoc effort to patch an existing system like in the case of the Web and P3P [Cranor et al., 2006].

Choice and Consent The EU Data Protection Directive of 1995 introduced for the first time a strong notion of consent in consumer privacy protection legislation. After its enactment, it was no longer sufficient to simply *notify* the subjects of data collection of such collection, but required the *unambiguous, free, specific and informed* consent before such collection occurs. Following the same

philosophy, Privacy by Design identifies this type of explicit consent as an important principle to consider during the design of systems. Note that here also, P3P assists achieving this principle in the web. This is because P3P enables browsers to act on behalf of the user, based on the user's preferences which can arguably be considered as some level of explicit, free and informed consent (or lack thereof, in case the privacy policy violates the user preferences).

Anonymity and Pseudonymity The principle of anonymity argues that since obtaining an explicit and informed consent is a difficult goal to achieve, then systems should consider collecting information in an anonymized and pseudonymized way so that consent will not be needed in that case.

Dissenting Opinion Even though this principle is seemingly innocent, we believe that it constitutes a dangerous slippery slope. It has been demonstrated once and over again that anonymized databases can be de-anonymized by multiple means including data fusion and/or exploiting special structures of these databases (e.g., sparsity) [Sweeney, 2002; Narayanan and Shmatikov, 2006, 2008]. To be fair, our understanding of the extent of this danger was only solidified after the publication of these principles in 2001; however, if anything, this demonstrates the difficulty of devising a fixed set of principles that will stand the test of time, because it is hard to foresee future dangers and risks.

Proximity and Locality The proximity principle states that in the absence of the ability to obtain explicit and informed consent, data collection should only occur whenever the owner of the ubiquitous artifact is in the proximity of that artifact. This should minimize the invasion of the privacy of individuals other than the owner of the device. For example, consider a sensor that collects heart rate in a non-contact manner [Poh et al., 2011, for example] for the purposes of assessing the risk of CHF. In this case, the sensor should only collect heart-rate data when the intended subject of such collection is in the proximity of the sensor. In contrast, the principle of locality attempts to alleviate the difficulty of identifying when the intended subject of the data collection is around, which could be a difficult problem for some sensors. In these cases, the locality principle dictates the collection of data only whenever the sensor is in the intended locality of data collection. For example, the non-contact heart rate monitor can alternatively only collect data whenever it is inside the emergency room of the hospital. Both of these principles are trying to deal with the difficulty that is inherent in obtaining informed and explicit consent from individuals in the setting of ubiquitous systems.

Partially Dissenting Opinion We believe that much like the principle of anonymity and pseudonymity, the inception of a whole new principle just for the sake of solving our inability to achieve another–more fundamental–principle, is a wrong and excessive approach that may end up backfiring. The main issue here is that an ad-hoc solution to a more fundamental problem is being declared a principle. While in this case, this ad-hoc solution is acceptable as a temporary solution, it does not address the fundamental issue at hand. In addition, to label this solution as a principle of equal value to the consent principle gives the impression that an adequate solution to the original fundamental issue has been reached. Instead, one should temporarily adopt the ad-hoc solution, but continue to work towards mitigating the gaps in achieving the fundamental principle. To give an analogy, if there is a pothole in the street (the fundamental issue), warning drivers with a sign that there is a pothole ahead is a temporary solution (the ad-hoc solution) until

the more fundamental issue is resolved (that is, closing the pothole). It would be almost comical to transform the real fundamental issue of closing the pothole into the question of how to present the warning to the drivers while accepting that this pothole will be there forever.

Adequate Security It is almost impossible to achieve an adequate level of privacy without an adequately secure infrastructure. However, it is important to note here that privacy and security are not interchangeable. You cannot replace one with another, but need both to operate in a coherent fashion in order to achieve the highest level of consumer protection possible. With this in mind, this principle is not necessarily easily achievable in general systems. Ubiquitous systems introduce a new set of security challenges given their form factor and computation power. Some of the more traditional and established storage and communication security technologies may not be applicable in ubiquitous computing. Therefore, special consideration to security needs to be given in the design phase of any system.

Access and Recourse This principle states that adequate access to the system details (not necessarily to the collected data) needs to be provided in order to be able to detect violations and enforce penalties or other remedies if necessary. This, in essence is trying to address the hard question of enforcement. In other words, in the design phase of a system, technology designers need to think about ways to allow audit of their systems by independent bodies and eventually enforcement of applicable policies, regulations and/or laws. Langheinrich argues that both of these topics belong more to the field of legal practice. He further argues that there is a need to revise or establish new laws and codes of conduct in order to address the special needs of ubiquitous computing. Nevertheless, part of the responsibility to enable enforcement, as discussed earlier, falls on the designers of such systems, and so this principle is a necessary one for the design of systems.

To summarize, Privacy by Design is a promising framework in which system designers incorporate privacy considerations into the design phase of their systems. We believe that this type of incorporation is necessary in order to effectively protect consumers' privacy. However, we do think that some fundamental gaps in achieving explicit and informed consent are present, and they need to be bridged. One potential way to achieve this is as follows. First, systems need to clearly communicate their policies in an unambiguous language that can be parsed by machines. This can be thought as being similar to the efforts carried out by the W3C P3P for websites. Once this is done, one can think of a universal privacy policy compliance system that reads all of these privacy policies, whether online or in the ubiquitous systems setting, and alerts the consumers whenever they are entering a space that doesn't comply with their preferences. The preferences need to be set once by the consumer in a manner that she or he understands (i.e., in a high-level language, for example: reject a policy that uses my home address for marketing purposes). To some extent, under this view, one can think of the world of privacy as if consumer were the ones writing their own privacy policies, and data collection agencies fall in different categories depending on their level of compliance with the consumer's privacy policy; based on these categories, consumers can decide whether or not to consent to the requested data collection. Even further, if data collectors can get access to the consumer's privacy policy, they may want to adapt their data collection pertaining to that consumer in order to comply with the consumer and not lose their business.

We believe that other gaps are present in protecting the privacy of consumers while communicating their data to curators or between curators. Specifically, we adopt a view of privacy that

indicates that privacy-breaches can still occur even if the data is secured in traditional means. This is possible through statistical reasoning about the data. Therefore, we feel that there is a need to introduce a new design principle that captures this threat. We elaborate more on this view in the next section.

3.3.2 Our Philosophical View on Privacy – The Inference Threat

Although Westin defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others," [Westin, 1968] Langheinrich argues that privacy is a notion that cannot be accurately defined [Langheinrich, 2001]. There is truth in both arguments, depending on what one wishes to distill from a definition. For instance, privacy is a social notion that gets refined with time, much like almost every other social notion; therefore, a single definition that will withstand the test of time will have to be generic in order to still apply after privacy concerns adapt with newer ways of life.

For actionable engineering purposes, it may be more useful to define a set of privacy goals that one wants to achieve, which is commonly referred to as a *privacy framework*. By identifying a privacy framework, system designers can then design their systems in a manner that is compliant to the adopted privacy framework. In this section, we identify a privacy framework that is relevant for the topic of this dissertation, predictive medicine.

By large we adopt the Privacy by Design framework, with the reservations described in Section 3.3.1. We argue that Privacy by Design as described lacks an important principle that is a product of modern techniques. In order to understand this proposed principle, let's conduct a thought experiment that is relevant to telemonitoring in particular and to predictive medicine in general.

A subject Bob is participating in an epidemiological study that aims at identifying risk factors for prostate cancer. The study utilizes telemonitoring technologies for the data collection. Among other health-related variables, the epidemiologists are interested in Bob's continuous respiratory rate. Bob consents in an informed fashion to this data collection. Some time later, it is discovered that respiratory rate can be used to accurately determine whether a person smokes marijuana or not. Bob would like to keep the information regarding his marijuana use private. Should Bob continue sharing continuous respiratory rate?

There are many possible answers to this nuanced question. For instance, in a setting where Bob trusts the medical researchers conducting the epidemiological study, it is sufficient for Bob to not consent for his data to be used for the purposes of predicting whether he smokes marijuana or not. In another scenario where Bob doesn't necessarily trust the doctors beyond his projected level of risk of harm (e.g., losing his job and/or increased insurance premium if the prediction came positive), Bob would most likely prefer to stop sharing the respiratory rate data with anyone. Regardless of the specific answer that each individual may take, this thought experiment demonstrates that there is a notion beyond the traditional *collection, use, retention and dissemination* of data. This notion is more nuanced, and has to do with what other pieces of information can be *inferred* from the collected data. From this, we propose the following principle of *Inference*.

Inference Data and information flows should be understood so that the privacy preferences of the consumer are taken into account when it comes to inference of undisclosed information from the collected data. Mechanisms need to be instilled in systems in order to audit and monitor such inferences and prevent or limit them in case they don't comply with the privacy preferences of

the subject of the inference. This principle also exemplifies the importance of solving the consent problem in a systematic way, because the notion of consent in the realm of inference should be updated to adapt to the new realities (e.g., consent to respiratory rate getting collected, but do not consent to using such data for the purposes of inferring whether Bob smokes marijuana or not). By considering this principle, it is easier to see why exchanging the roles of who dictates and who follows privacy policies has the potential to provide a stable data ecosystem.

The principle of Inference can also be viewed by contrasting privacy to traditional security. In traditional encryption approaches to maintaining privacy, it is often implicitly assumed that the data themselves *are* the private information. However, in some scenarios, the data *can be used* to infer certain private information about the subjects from the given data. For example, respiration rate by itself might not be considered private information. However, if the data from the collected respiration rate are used to infer whether the individual is a marijuana smoker or not, they become sensitive information. One can argue that because the information about whether someone smokes marijuana is private, the respiration rate data become private *by extension*.

Finally, we advocate the guideline that the specific privacy concerns need to be studied before the deployment of any system. In this context, we propose a guideline for privacy-aware system design that helps achieve that, the *2.0 from Scratch* guideline (note that we don't call it a principle but merely a guideline).

2.0 from Scratch This guideline advises system designers to implement a first version of their system for the purposes of privacy and acceptability studies, similar to those presented in Section 3.2. Based on the findings of such studies, it is advised to start the second major version of the system from scratch. There are similar guidelines in traditional systems design for building large robust systems. In those guidelines, software engineers are advised to build the second version of a large system from scratch, incorporating all the lessons learned from the first version.

With the views set forth in this section, we developed a framework for information disclosure that protects consumers against inference by unauthorized entities in a MITM attack setting. We introduce this framework in the following section. Note that even though this framework is applicable to any passive MITM inference-based attacks, we introduce it in a language that is consistent with telemonitoring for coherency.

3.4 Private Disclosure of Information (PDI)

3.4.1 Introduction

Telemonitoring can be considered a health-related ubiquitous systems, which implies that it comes with all of the privacy concerns that we discussed thus far. Not only is privacy-awareness an important ethical consideration to make during the design of telemonitoring systems, but it has serious implications on the reliability of the collected data. For instance, Warner argued, in 1965, that individuals would be unwilling to share their data, or simply share false information (if possible), in the absence of privacy guarantees [Warner, 1965]. Therefore, lack of privacy protection could result in bias due to the choice of some individuals to not share their information, or simply because of the accuracy of the collected information, which in turn hinders our ability to draw

conclusions from such data.[2] As a result, we need to understand the privacy threats specific to telemonitoring, and design mechanisms to protect the privacy of the users of such technology. In addition, such an effort would greatly promote the adoption of telemonitoring.

Let us start by identifying the life-cycle stages of information curated and handled in telemonitoring systems. These stages include *i*) the collection and disclosure of health-related data by the users; *ii*) the processing and analysis of the data by the telemonitoring servers; *iii*) the publishing of results based on the data, which may include publishing a subset of the dataset, statistics of it, a sanitized version of it, or a combination of those.

In this section, we focus on defending against inference threats that are present in the first stage of the information's life-cycle described above. Concretely, we aim to reduce the ability of an eavesdropper to infer certain privacy-sensitive information about the patient, such as the diagnosis, from the information that the patient discloses to their doctors, through telemonitoring. This threat is usually called an *inference attack*. As such, we adopt the *Inference* point of view described in Section 3.3.2 and design a telemonitoring system that conforms with the *Inference* principle presented therein. The framework that is designed for this purpose, and presented in this section, is called Private Disclosure of Information (PDI) [Aranki and Bajcsy, 2015].

By focusing on the *Inference* design principle, we aim to sanitize the transmitted (disclosed) information in a way that leaks as little as possible about other private (undisclosed) pieces of information to an eavesdropper. In summary, the objective is to sanitize the disclosed data in order to hide other private pieces of information, which can be *infered* from these data. As discussed earlier in this chapter, privacy and security are not interchangeable. As Sweeney wrote, "computer security is not privacy protection" [Sweeney, 2002]. The converse is also true, privacy does not replace security. This is an important observation; therefore, we emphasize that PDI is complementary to, not a replacement of, classical security approaches. For instance, encryption can be applied to the PDI-sanitized data.

3.4.2 Problem Formulation

Notation

Until the end of this chapter, we use the notation convention described below. We denote a random variable by a capital letter (say X), and denote a realization of a random variable by the corresponding small letter (i.e., x). In order to reduce notational overload, we infer which probability density (mass) function we refer to by the letter(s) used inside. For instance, for the conditional density (mass) function of Y given X, we simply write $p(y|x)$, and for the marginal density (mass) function of X we write $p(x)$.

Setting and Threat Model

We introduce the setting and the threat model through the following scenario. A patient, Bob, is diagnosed by his physician, Alice, with a health condition c. Alice, in turn, would like to get regular updates about Bob's health status. As such, she utilizes telemonitoring in order to get regular updates about Bob's symptoms, vital signs and other health-related variables. For the purposes of the discussion on PDI, the objective is not to perform a diagnosis, but to remotely monitor Bob's health *after* a diagnosis is obtained (e.g., for prognosis purposes).

[2] We elaborate, in detail, on the dangers of bias in Chapter 4.

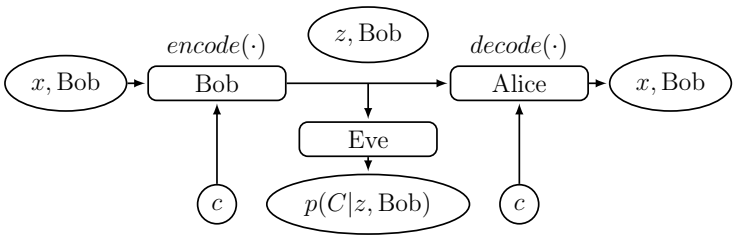


Figure 3.6: The PDI threat model: a statistical inference attack by a passive eavesdropper.

The data, x, that Bob sends through telemonitoring, is sensitive from a privacy point of view, and subjects Bob to statistical inference threats. For instance, an eavesdropper, Eve, can use x to infer Bob's diagnosis, c, through statistical reasoning. As a result, we aim to formulate an information disclosure process, in which Bob is protected from Eve's inference attack. Moreover, the process should be compatible with security techniques that protect the disclosed message x itself (e.g., cryptography). We achieve this protection by encoding the data, x, that Bob wishes to disclose, and sending that instead. The encoded information, which Bob actually sends to Alice through telemonitoring, is denoted by z (i.e., the sanitized version of x).

Consequently, the threat model we are considering, depicted in Figure 3.6, is as follows. Bob wishes to disclose his health-related information x to Alice. Alice already knows the diagnosis, c, of Bob. Eve, who does not originally know Bob's diagnosis c, wishes to learn it from Bob's disclosed information through telemonitoring (which does not include c itself). To protect himself from this threat, Bob discloses a sanitized version of the information, z.

As such, Eve treats the diagnosis as a random variable, C, and reasons about it in a probabilistic manner. Concretely, Eve updates her belief regarding Bob's diagnosis after observing z. Namely, Eve updates $p(C|Bob)$ to $p(C|Bob, z)$. Note that in this setting, we are assuming that the identity of the sender is known to Alice and Eve. Hence, the threat model is an inference attack by a passive eavesdropper.

The objective is therefore to find a way to encode x to z such that Eve's ability to gain information about Bob's diagnosis, c, from z is minimized. In contrast, cryptography aims to limit Eve's ability to infer the original message x–not c–from z. Using the notation we introduced, cryptography aims to limit Eve's ability to update her belief $P(X|Bob)$ to $p(X|Bob, z)$, where X is a random variable describing the original information Bob wants to send. We note that these two goals are related, but are also different. Several inference attacks, based on encrypted messages, were demonstrated in the literature [see White et al., 2011; Miller et al., 2014, for example].

To summarize, let s be the identifier of the patient disclosing information using telemonitoring (Bob in the scenario above). We aim to formulate a process of information disclosure that satisfies the following premises:

DECODING The physician (Alice) is able to fully utilize the disclosed information z by being able to retrieve the original message x from z; and

HIDING CLASS The eavesdropper's (Eve's) ability to perform statistical *inferences* on c (given s), based on the disclosed information, z, is minimized.

Definitions

Formally, we denote the set of identifiers of the patients who use the telemonitoring system by $\mathbb{S}$. Moreover, $\mathbb{I}$ denotes the information space of health data. The set of all diagnoses that patients

wish to keep private is denoted by $\mathbb{C}$.

Similarly, we denote the random variables of the patient's identity by S and the random variable of the (private) diagnosis by C. Moreover, we denote the random variable of the piece of information that the patient would like to disclose by X. Finally, we denote the random variable of the disclosed piece of information, after encoding, by Z. Z is often referred to as the sanitized information.

Consider the following form of encoding schemes.

Definition 3.1. *A Privacy Mapping Function (PMF) is a function $R : \mathbb{C} \to \mathbb{I}^{\mathbb{I}}$, where $\mathbb{I}^{\mathbb{I}}$ is the set of injective functions $\mathbb{I} \to \mathbb{I}$.*

A PMF, R, works as follows. Given a diagnosis $c \in \mathbb{C}$, $[R(c)](\cdot)$ is an injective encoding function that sanitizes information generated by patients with diagnosis c. Concretely, a patient s with diagnosis c sanitizes her or his data by calculating $z = [R(c)](x)$ (encode$(\cdot)$ from Figure 3.6).

Since Alice alredy knows the diagnosis, c, of the patient s, she can indeed retrieve back x from z by calculating $x = [R(c)]^{l}(z)$ (decode$(\cdot)$ from Figure 3.6). Here, $[R(c)]^{l}(\cdot)$ is a left inverse of $[R(c)](\cdot)$.[3] This means that the premise (**DECODING**) is satisfied by construction.

From the discussion above, we conclude that the random variables Z, X and C are related by $Z = [R(C)](X)$. The statistical graphical model relating the random variables is depicted in Figure 3.7.

Illustrating Example

We demonstrate the definitions and concepts presented thus far by the following example. In this example, we simulate a telemonitoring system that tracks the Body Mass Index (BMI) and weight of subjects who are 19 years of age or younger.[4] The collection of all patients constitutes the set $\mathbb{S}$. In our example, the information space is defined as $\mathbb{I} \triangleq \{(bmi, w) \in \mathbb{R}^2\}$, where bmi is the patient's BMI $[\frac{\text{kg}}{\text{m}^2}]$ and w is the patient's weight [kg].

In this example, we consider the following diagnoses $\mathbb{C} \triangleq \{UW, HW, OW, OB\}$ denoting underweight, health weight, overweight and obese, respectively. The Center for Disease Control and Prevention (CDC) defines these as weight categories, where each child or teen is classified based on her or his BMI percentile among the same gender and age group. These percentiles are described in Table 3.1.

Based on the definitions in Table 3.1, Eve's prior belief of diagnoses for any patient $s \in \mathbb{S}$ (from our monitored subjects), assuming she doesn't have extra information about any particular patient, are

$$p(C = UW|s) = 0.05, p(C = HW|s) = 0.8, p(C = OW|s) = 0.1 \text{ and } p(C = OB|s) = 0.05.$$

Intuitively, Eve's ability to infer the subject's weight category from the the information $x \in \mathbb{I}$ is not perfect (even if she observes it plainly), since the age and gender are not disclosed.

However, Eve may gain some knowledge regarding the diagnosis by observing the communicated messages. For example, assume that a subject s discloses the message $x = (50\,\frac{\text{kg}}{\text{m}^2}, 120\,\text{kg}) \in \mathbb{I}$ without encoding (i.e., $z = x$). This datapoint implies that s weighs 120 kg and measures about 155 cm tall.[5]

[3] We say that $g : D_2 \to D_1$ is a left inverse of a function $f : D_1 \to D_2$ if for all $x \in D_1$ we have $g(f(x)) = x$.

[4] BMI is a measure of relative weight based on an individual's mass and height. Defined as $bmi \triangleq \frac{mass[\text{kg}]}{height^2[\text{m}^2]}$.

[5] The height can be calculated as $\sqrt{\frac{mass}{BMI}} = \sqrt{\frac{120}{50}} \approx 1.55\,\text{m} = 155\,\text{cm}$.

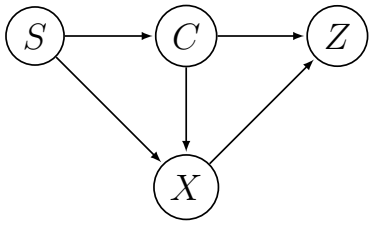


Figure 3.7: The statistical graphical model of PDI. S and C are the patient's identity and diagnosis, respectively; X is the information intended for disclosure; and Z is the sanitized (encoded) information that gets disclosed.

Weight Category	**BMI Percentile Range**
Underweight	$BMI < 5\%$
Healthy Weight	$5\% \leq BMI < 85\%$
Overweight	$85\% \leq BMI < 95\%$
Obese	$95\% \leq BMI$

Table 3.1: The different weight categories defined by the corresponding BMI percentiles within the same age and gender group for individuals of age 19 or less. This definition is consistent with the one provided by the CDC.

Even without observing the subject's gender or age, Eve's posterior belief $p(C = OB|s, Z = (50, 120))$ will increase (compared to the prior 0.05). Intuitively speaking, this is because it is more likely that such BMI and weight correspond to a subject diagnosed with obesity than it is the case that they correspond to a subject with an unknown diagnosis.

Formally, Eve's probabilistic reasoning relies on Bayes' rule

$$p(c|s, Z = (50, 120)) = \frac{p(Z = (50, 120)|c, s)}{p(Z = (50, 120)|s)} p(c|s)$$

for the different diagnoses $c \in \mathbb{C}$. Since $p(Z = (50, 120)|C = OB, s)$ is higher than $p(Z = (50, 120)|s)$, the ratio $\frac{p(Z=(50,120)|C=OB,s)}{p(Z=(50,120)|s)}$ is higher than 1 and therefore the posterior $p(C = OB|s, Z = (50, 120))$ will be higher than the prior $p(C = OB|s) = 0.05$. Alternatively, if we encode the message $x = (50, 120) \in \mathbb{I}$ to $\hat{z}$ in a way that yields

$$\frac{p(\hat{z}|c, s)}{p(\hat{z}|s)} \approx 1, \forall c \in \mathbb{C},$$

we limit Eve's ability to update her belief about the diagnosis of patient s because her posterior belief will be approximately equal to her prior belief. PDI's objective is to generalize this intuition for every message, diagnosis and subject.

Satisfying the HIDING CLASS Premise

Since the premise (**DECODING**) is guaranteed by construction, we now move to discuss the (**HIDING CLASS**) premise. Generally speaking, in order to fully describe the statistical graphical model depicted in Figure 3.7, we have to provide the following distributions.

- $p(s)$, the distribution of patients transmitting messages in the telemonitoring system.
- $p(c|s)$, Eve's prior of patients' diagnoses, based on auxiliary knowledge.
- $p(x|c, s)$, the generative model of health data given a diagnosis and a patient.

Based on the definition of PMF, $p(z|x, c)$ be simply described by

$$\mathbb{P}\left(Z = z|X = x, C = c\right) = \begin{cases} 1 & \text{if } z = \left[R(c)\right](x) \\ 0 & \text{otherwise} \end{cases}$$

for all $z, x \in \mathbb{I}$ and $c \in \mathbb{C}$.

Given the statistical model described in Figure 3.7, we now turn to generalize the intuition we built in the illustrating example. To generalize that intuition, we aim to find a PMF, R, that makes the posterior belief $p(C|s,z)$ as close as possible to the prior belief $p(C|s)$ for all $s \in \mathbb{S}$ and $z \in \mathbb{I}$. This can be done through a choice of R that makes the ratio $\frac{p(z|c,s)}{p(z|s)}$ as close to 1 as possible for all $s \in \mathbb{S}$, $c \in \mathbb{C}$ and $z \in \mathbb{I}$.

In order to achieve this, we have to introduce some mathematical tools and build to that result. In that journey, we first introduce the notion of Conditional Mutual Information.

Definition 3.2. *[Cover and Thomas, 2006, c.f. Definition 8.49] Let X, Y and Z be random variables. The* conditional mutual information of X and Y given Z, *$I(X,Y|Z)$, is defined as*

$$I(X,Y|Z) \triangleq \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right]$$

We will use conditional mutual information to measure the quality of our PMF, R, by considering the value of $I(Z,C|S;R)$. This notation implies that the conditional mutual information, in our case, is a function of R. Intuitively, $I(Z,C|S;R)$ is a measure of the expected amount of information Z carries about C (and vice versa). If the log base in Definition 3.2 is 2, then the (conditional) mutual information outputs a measure in bits. As such, it is desirable to find a PMF, R, that minimizes $I(Z,C|S;R)$. In order to see this, we present the following known result that ties conditional mutual information to conditional independence of random variables.

Lemma 3.1. *[Cover and Thomas, 2006, c.f. Corollary 2.92; c.f. Theorem 8.6.1] $I(Z,C|S;R) \geq 0$ for any PMF, R. Furthermore, $I(Z,C|S;R) = 0$ if and only if Z and C are conditionally independent given S; using the PMF, R.*

Jiao et al. axiomatically justified the use of (conditional) mutual information as a measure of privacy for side information, a result that is applicable to our setting [Jiao et al., 2014]. Tying this intuition together and formalizing it mathematically, we attempt to satisfy (**HIDING CLASS**) by finding a PMF that minimizes the conditional mutual information between Z and C, given S as follows.

$$R^* = \underset{R \text{ is a PMF}}{\arg\min}\, I(Z,C|S;R) \tag{3.1}$$

Finally, once this PMF is found (and can be made public), we describe the process of Private Disclosure of Information (PDI) as follows:

Sending In order for a patient, diagnosed with condition $c \in \mathbb{C}$, to disclose a piece of information $x \in \mathbb{I}$, she or he first sanitizes x by applying $z \leftarrow [R(c)](x)$ and then sends z (or some encrypted version of it).

Receiving In order to interpret a message $z \in \mathbb{I}$, the intended recipient (who knows the diagnosis c of the patient), applies $x \leftarrow [R(c)]^l(z)$, where $[R(c)]^l(\cdot)$ is a left inverse of $[R(c)](\cdot)$.

Two questions now arise. First, why does minimizing the conditional mutual information measure as described in Equation (3.1) protect against the inference attack described in our threat model? Second, given the appropriate data, how do we compute a solution to the problem described in Equation (3.1), noting that it is not a convex optimization problem? In the next two

sections, we address these two questions in detail. In Section 3.4.3 we formally tie Equation (3.1) to our main objective, and present a few theoretical results in that regard. Afterwards, in Section 3.4.4, we present a MATLAB[6] toolbox was developed by the Berkeley Telemonitoring project that implements the learning problem described in Equation (3.1), from data [Aranki and Bajcsy, 2016].

3.4.3 Further Analysis and PDI Properties

First, let us we relate the value of the objective function in Equation (3.1) to what we are trying to defend against in the following lemma. Concretely, we will relate the value of conditional mutual information to the ability to perform Bayesian updates, the backbone of Bayesian inference.

Lemma 3.2. *If a PMF, R, yields $I(Z, C|S; R) = 0$ then Bayesian updates for the belief of C based on observing Z, given S, are prevented for the eavesdropper. Formally, $p(c|s, z; R) = p(c|s)$ for all $c \in \mathbb{C}, s \in \mathbb{S}$ and $z \in \mathbb{I}$.*

Proof. From Lemma 3.1 we know that by using this PMF, R, we get that Z is conditionally independent of C given S which means $p(c|s, z; R) = p(c|s)$ for all $c \in \mathbb{C}, s \in \mathbb{S}$ and $z \in \mathbb{I}$. This reads that the posterior of the diagnosis is equal to the prior of the diagnosis that the eavesdropper already possesses, which is precisely what needs to be shown. Therefore, the disclosure of Z does not change the eavesdropper's belief regarding diagnosis C given the patient identifier S. □

Let us explain why such a PMF is desirable by intuition. If a PMF, R, satisfies the hypothesis of Lemma 3.2, then

$$p(c|s) \overbrace{=}^{\text{Lemma 3.2}} p(c|s, z; R) \overbrace{=}^{\text{Bayes' rule}} \frac{p(z|c, s)}{p(z|s)} p(c|s), \forall c \in \mathbb{C}, s \in \mathbb{S}, z \in \mathbb{I}$$

$$\implies \frac{p(z|c, s)}{p(z|s)} = 1, \text{ whenever } p(c|s) \neq 0 \tag{3.2}$$

First, this says that whatever prior belief the eavesdropper, Eve, possesses that subject s has diagnosis c, will remain the same after observing the sanitized message z that s sent. In other words, the the information contained in the sanitized message z that s discloses will not change how Eve perceives the odds of s having diagnosis c (i.e., Eve did not learn anything new regarding the diagnosis of s by observing the disclosed message z). Moreover, recalling our intuition from the illustrating example in Sections 3.4.2 and 3.4.2, it is desirable to have a ratio $\frac{p(z|c,s)}{p(z|s)}$ as close to 1 as possible, which this lemma ensures.[7] However, Lemma 3.2 is merely a statement, which is *conditioned* on the existence of such a PMF, R, and does *not* speak to its existence.

Therefore, the next question that we need to ask is whether a PMF, R, satisfying $I(Z, C|S; R) = 0$ is ever attainable. There are two reasons to be suspicious of the existence of such PMF, R. First, if such a PMF, R, exists, then it means that by knowing S (which is always attached to the message), Z provides no extra information to inferring C to an eavesdropper, which sounds surprising. Second, as discussed earlier, there is generally a trade-off between information utility and privacy, where optimal privacy is usually *only* attained at the cost of no (or very little)

[6] `https://www.mathworks.com/products/matlab/`

[7] When the prior belief of Eve $p(c|s) = 0$, no information will "change her mind". As in, she will always believe that s doesn't have diagnosis c regardless of the disclosed information z she observes. This can easily be seen from applying Bayes' rule: if $p(c|s) = 0$ then $p(c|s, z; R) = \frac{p(z|c,s)}{p(z|s)} p(c|s) = 0$.

utility [Dwork, 2006]. However, in our case, the utility of the information Z to the doctor is always fully preserved, irrespective of the choice of R (as long as it is a PMF), since $[R(c)](\cdot)$ is injective for all $c \in \mathbb{C}$, which allows the doctor to always decode z back to the original message x. From this, it follows that the scenario of perfect privacy seems to be unattainable.[8] All that said, if such a PMF, R, exists, it would assure optimality of Equation (3.1), so it would be desirable to show its existence, if it does exist. Fortunately (and somewhat unintuitively), such a PMF can be attained as shown in the following sequence of results.

Lemma 3.3. *If there exists a function $f(z,s)$ such that $p(z|c,s) = f(z,s)$ for all $c \in \mathbb{C}, z \in \mathbb{I}$ and $s \in \mathbb{S}$ then $p(z|s) \equiv f(z,s)$.*

Proof.

$$\begin{aligned}
p(z|s) &\overset{\text{marginalize}}{=} \sum_{c\in\mathbb{C}} p(z,c|s) \\
&\overset{\substack{\text{conditional}\\ \text{probability}}}{=} \sum_{c\in\mathbb{C}} p(z|s,c)\cdot p(c|s) \\
&\overset{\text{hypothesis}}{=} \sum_{c\in\mathbb{C}} f(z,s)\cdot p(c|s) \\
&\overset{\text{factor out}}{=} f(z,s)\cdot\sum_{c\in\mathbb{C}} p(c|s) \\
&\overset{\substack{\text{probability}\\ \text{axioms}}}{=} f(z,s)
\end{aligned}$$

□

Using Lemma 3.3, we prove the following theorem, which is a sufficient condition for the optimality of Equation (3.1). This theorem, comes intuitively from our discussion following Equation (3.2).

Theorem 3.1. *If there exists a function $f(z,s)$ such that $p(z|c,s) = f(z,s)$ for all $c \in \mathbb{C}, z \in \mathbb{I}$ and $s \in \mathbb{S}$, then $D_{KL}(p(c|z,s)||p(c|s)) = 0$ for all $z \in \mathbb{I}$ and $s \in \mathbb{S}$.*[9]

Proof. Since $p(z|c,s) = f(z,s)$ then using Lemma 3.3 we know that $p(z|s) \equiv f(z,s)$. Therefore, for any $z \in \mathbb{I}$ and $s \in \mathbb{S}$ such that $f(z,s) = p(z|c,s) = p(z|s) \neq 0$ we get

$$\begin{aligned}
\frac{p(c|z,s)}{p(c|s)} &\overset{\substack{\text{Bayes' rule on}\\ p(c|z,s)}}{=} \frac{p(z|c,s)\cdot p(c|s)}{p(c|s)\cdot p(z|s)} \\
&\overset{\text{Cancel terms}}{=} \frac{p(z|c,s)}{p(z|s)} \\
&\overset{\text{Lemma 3.3}}{=} \frac{f(z,s)}{f(z,s)} \\
&= 1
\end{aligned}$$

This implies, by the definition of Kullback-Leibler divergence, that $D_{KL}(p(c|z,s)||p(c|s)) = 0$, as requested. □

[8] Since we are treating the setting of Bayesian inference, we consider "perfect privacy" to be that the eavesdropper's belief about C given S doesn't change after observing Z, in line with Lemma 3.2.

[9] $D_{KL}(p||q)$ is the Kullback-Leibler divergence from q to p.

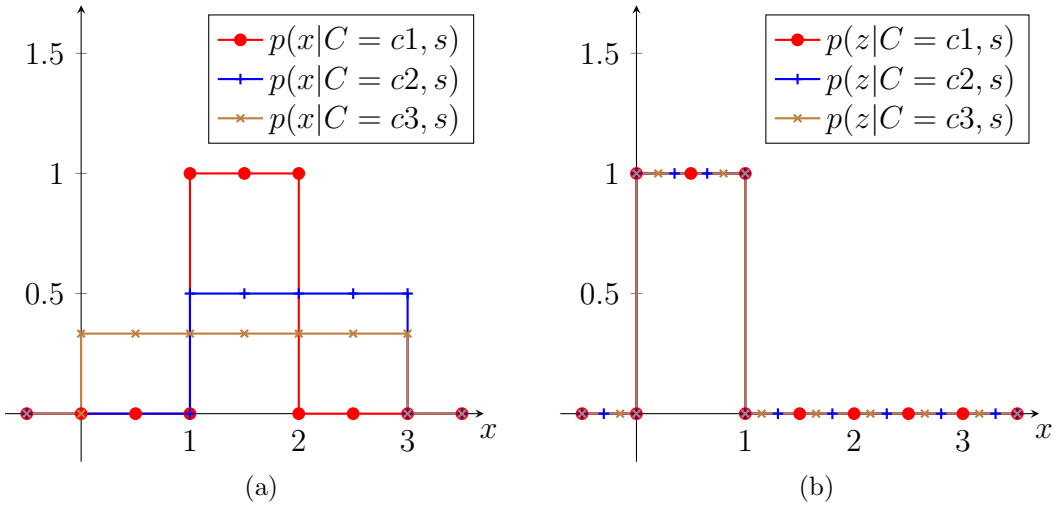


Figure 3.8: Demonstrating the "folding" intuition. (a) The data distribution, before sanitizing; and (b) the sanitized data distribution ("folded").

Corollary 3.1. *If a PMF, R, achieves $p(z|c,s) = f(z,s)$ for some function $f(z,s)$, for all $c \in \mathbb{C}, z \in \mathbb{I}$ and $s \in \mathbb{S}$, then R is the optimal solution to Equation* (3.1).

Proof. The result follows from Theorem 3.1 and the fact that

$$I(Z, C|S; R) = \mathbb{E}_{p(z,s)}[D_{KL}(p(c|z, s; R)||p(c|s; R))].$$

□

Lemma 3.3, Theorem 3.1, and Corollary 3.1 enforce and formalize the intuition that we developed in Section 3.4.2. The intuition is that the closer $p(z|c,s)$ is to $p(z|s)$, the more privacy-preserving the PMF, R, is. An alternative and complementary intuition is that the aim is to devise a PMF, R, that statistically "folds" the information Z for all diagnosis such that they look the same irrespective of $c \in \mathbb{C}$. That is, Corollary 3.1 dictates that if the distributions $p(z|c,s;R) = f(z,s)$ are the same, irrespective of c (the "folding"), then R is an optimal PMF.

To illustrate the "folding" intuition, consider the toy example depicted in Figure 3.8. In this toy example, we are considering one subject only $\mathbb{S} = \{s\}$, the information space is the reals $\mathbb{I} = \mathbb{R}$, and $\mathbb{C} = \{c_1, c_2, c_3\}$. In Figure 3.8a, the distributions of the raw (not sanitized) data are depicted, for the subject, under the assumption that the subject is diagnosed with each one of the three diagnoses (since the eavesdropper doesn't know the subject's diagnosis). For example, if the subject transmits the datapoint $z = x = \frac{1}{2}$ (i.e., without sanitizing) then the eavesdropper can know for sure that the subject is diagnosed with c_3. However, if we can find a PMF, R, that sanitizes the data so that their distributions for all 3 diagnoses look the same ("folding"), as depicted in Figure 3.8b, then we can completely defend against inference attacks. This is because, if the subject now transmits the sanitized datapoint $z = 0.5$ that was sanitized using R, the eavesdropper will not become any more or less confident about his prior belief of the diagnosis of s, because the class that generated this datapoint is statistically indistinguishable from the rest. This intuition will be further demonstrated and discussed more thoroughly with a real-world example in Section 3.4.5.

Note that Theorem 3.1 is *irrespective* of the model of $p(c|s)$ (and $p(s)$). This is a very important observation since it means that in cases where a PMF, R, satisfies the condition of the theorem,

modeling the eavesdropper's prior knowledge about patients' diagnoses is not needed. This is important because it is often hard to model the eavesdropper's auxiliary knowledge, and that sometimes different eavesdroppers have different auxiliary knowledges. Furthermore, such PMF achieves perfect privacy against any eavesdropper, regardless of her auxiliary knowledge $p(c|s)$ (or $p(s)$). In the following theorems we provide examples of using Theorem 3.1 that also serve as cases proving that such PMFs, achieving perfect privacy, are attainable.

Theorem 3.2. *If $X|C=c,S=s \sim N(\mu_c,\Sigma_c)$ (Normal distribution) for every $c \in \mathbb{C}$ and $s \in \mathbb{S}$, then $[R(c)](x) = \Sigma_c^{-\frac{1}{2}} \cdot (x-\mu_c)$ is an optimal solution to Equation (3.1), and achieves perfect privacy.*

Proof. It is easy to verify that $Z|C=c,S=s \sim N(\vec{0},I)$ for every $s \in \mathbb{S}$ and $c \in \mathbb{C}$, where $\vec{0}$ is the origin in the information space (vector of zeros) and I is the identity matrix (of the appropriate dimensions). This means that $p(z|c,s) \equiv f(z,s)$ (not a function of c). By using Corollary 3.1, we therefore know that R is the optimal solution to Equation (3.1), and achieves perfect privacy. □

This is proof, by construction, that such PMFs are attainable. Similarly, we show the following results.

Theorem 3.3. *If $X|C=c,S=s \sim Exp(\lambda_c)$ (Exponential distribution) for every $c \in \mathbb{C}$ and $s \in \mathbb{S}$, then $[R(c)](x) = \lambda_c x$ is an optimal solution to Equation (3.1), and achieves perfect privacy.*

Proof. It is easy to verify that $Z|C=c,S=s \sim Exp(1)$ for every $s \in \mathbb{S}$ and $c \in \mathbb{C}$. This means that $p(Z=z|C=c,S=s) \equiv f(z,s)$ (not a function of c). By using Corollary 3.1, we therefore know that R is the optimal solution to Equation (3.1), and achieves perfect privacy. □

Theorem 3.4. *If $X|C=c,S=s \sim Gamma(k,\theta_c)$ (Gamma distribution with shape and scale parameters) for every $c \in \mathbb{C}$ and $s \in \mathbb{S}$, then $[R(c)](x) = \frac{x}{\theta_c}$ is an optimal solution to Equation (3.1), and achieves perfect privacy.*

Proof. It is easy to verify that $Z|C=c,S=s \sim Gamma(k,1)$ for every $s \in \mathbb{S}$ and $c \in \mathbb{C}$. This means that $p(Z=z|C=c,S=s) \equiv f(z,s)$ (not a function of c). By using Corollary 3.1, we therefore know that R is the optimal solution to Equation (3.1), and achieves perfect privacy. □

Theorem 3.5. *If $X|C=c,S=s \sim U(a_c,b_c)$ (Continuous Uniform distribution) for every $c \in \mathbb{C}$ and $s \in \mathbb{S}$, then $[R(c)](x) = \frac{x-a_c}{b_c-a_c}$ is an optimal solution to Equation (3.1), and achieves perfect privacy.*[10]

Proof. It is easy to verify that $Z|C=c,S=s \sim U(0,1)$ for every $s \in \mathbb{S}$ and $c \in \mathbb{C}$. This means that $p(Z=z|C=c,S=s) \equiv f(z,s)$ (not a function of c). By using Corollary 3.1, we therefore know that R is the optimal solution to Equation (3.1), and achieves perfect privacy. □

The intuition set forth here is further empirically demonstrated in Section 3.4.5 where the data from each diagnosis, after sanitization, are mapped onto a distribution that is similar to that of the data from the other diagnoses (Figure 3.10) to limit inference attacks.

[10]Note that the example demonstrated in Figure 3.8 is a direct application of Theorem 3.5.

3.4.4 Learning the Privacy Mapping Function

As mentioned earlier, the problem from Equation (3.1) is not a convex optimization problem. Moreover, it is often the case that the models for $p(x|c,s)$ and $p(c|s)$ are not explicit. Instead, data can be available for x and c. In this section, we discuss our implementation of learning a PMF from such data. Our implementation takes the form of a MATLAB toolbox that is publicly available in open source [Aranki and Bajcsy, 2016].

Given a dataset $D = \left\{(x_i, c_i)_{i=1}^N | x_i \in \mathbb{I}, c_i \in \mathbb{C}\right\}$ with the data x_i corresponding to a patient with diagnosis c_i, the problem at hand is to learn an optimal PMF, R, according to Equation (3.1), from D. We first note, that our treatment in this implementation doesn't include the subject identifiers s_i. This simplifying assumption of ignoring the modeling of the random variable S for the learning problem has the following implications on the full model from Figure 3.7.

First, this simplifying assumption implies that S is treated as a uniform random variable. That is, $p(s) = \frac{1}{|\mathbb{S}|}$ for all $s \in \mathbb{S}$. Informally, this means that the subjects are equally likely to disclose information using the telemonitoring system.

The second implication is that S is treated as an independent random variable of C. That, in turn, implies $p(c|s) = p(c)$ for all $s \in \mathbb{S}, c \in \mathbb{C}$. Informally, this means that Eve has no special prior knowledge about the diagnosis of any specific subject and that her prior belief about the subjects' diagnoses is equal for all subjects. We note here that for the cases of perfect privacy discussed in Theorem 3.1 and Corollary 3.1, this implication does not degrade the quality of the learned PMF R. This is because, per the discussion in Section 3.4.3, the solution for R in such cases is irrespective of Eve's prior belief $p(c|s)$. Further study is needed to assess and quantify the privacy-degradation incurred by this simplifying assumption in cases of imperfect privacy.

Third, this assumption implies that, given C, X and S become independent. That is, $p(x|c,s) = p(x|c)$ for all $s \in \mathbb{S}, c \in \mathbb{C}$ and $x \in \mathbb{I}$. Informally, this means that the generative distribution of the disclosed telemonitoring data, X, is not a function of the subject, once the diagnosis is known. In our example, this means that once we know the BMI category of the subject, the distribution of the subject's weight and BMI becomes known, irrespective of which subject is being reasoned about.

Finally, this assumption implies that $I(Z, C|S; R) = I(Z, C; R)$, which simplifies the learning problem of R.

We now turn to discuss the search space of Equation (3.1). This is the space of all PMFs $\mathcal{F} = \left\{R : \mathbb{C} \to \mathbb{I}^{\mathbb{I}}\right\}$. Note that performing computation over this space is intractable. Therefore, for our implementation, we limit the search space to a smaller subset of $\mathcal{F}$ that can be characterized parametrically. In the toolbox, the user can define any parametric subset of $\mathcal{F}$ for the search space. Once this parametric characterization is provided, Equation (3.1) can be rewritten using the corresponding parameter space, Θ, as follows:

$$\theta^* = \arg\min_{\theta \in \Theta} I(Z, C; R(\cdot; \theta)) \tag{3.3}$$

For example, the following space is a parametric space of all PMFs, R, that yield injective affine transformations for all $c \in \mathbb{C}$.

$$\mathcal{F}_{\text{affine}} \triangleq \left\{ R : \mathbb{C} \to \mathbb{I}^{\mathbb{I}} \,\middle|\, \substack{[R(c)](x) = A_c \cdot (x - b_c), \\ A_c \in \mathbb{R}^{n \times n}, A_c \text{ is invertible}, b_c \in \mathbb{R}^n, \forall c \in \mathbb{C}} \right\} \tag{3.4}$$

where n is the number of elements in the vector x. This yields the following parameter space

$$\Theta_{\text{affine}} \triangleq \left\{(A_c, b_c)_{c \in \mathbb{C}} | A_c \in \mathbb{R}^{n \times n}, A_c \text{ is invertible}, b_c \in \mathbb{R}^n, \forall c \in \mathbb{C}\right\}$$

Algorithm 3.1 calcMI $(R, D, p(c))$ – the objective function

Input: $R : \mathbb{C} \to \mathbb{I}^{\mathbb{I}}$: PMF.
Input: $D = \{(x_i, c_i)_i\}$: the input dataset.
Input: $p(c)$: the prior belief of the random variable C.
Output: $I(Z, C; R)$

1: **for** $c \in \mathbb{C}$ **do**
2: $\quad \mathbb{X}_c \leftarrow \{x_i | c_i = c\}$
3: $\quad \mathbb{Z}_c \leftarrow \{[R(c)](x) \,| x \in \mathbb{X}_c\}$
4: $\quad h_c \leftarrow \text{hist}(\mathbb{Z}_c)$
5: $\quad p(z|c) \leftarrow \frac{h_c}{|\mathbb{Z}_c|}$
6: **end for**
7: $p(z) \leftarrow \sum_{c \in \mathbb{C}} p(z|c) \cdot p(c)$
8: **return** $\sum_{z \in \mathbb{I}, c \in \mathbb{C}} p(z|c) \cdot p(c) \cdot \left[\log \frac{p(z|c)}{p(z)}\right]$

We are now ready to describe the inner workings of our learning procedure. Given a dataset D, the toolbox models $p(c)$ non-parametrically using a histogram. Using this histogram, and provided a parameter $\theta \in \Theta$ (which defines a realization of a PMF, $R(\cdot;\theta)$), calculating $I(Z, C, R(\cdot;\theta))$ is straightforward. This can be done by setting $z_i = [R(c_i)](x_i)$ for all $(x_i, c_i) \in D$, and using (z_i, c_i) to construct high dimensional histograms modeling $p(z|c)$ and $p(z)$ non-parametrically. Using this histogram and $p(c)$, the value of the mutual information $I(Z, C; R)$ can be computed as follows

$$\begin{aligned} I(Z, C; R(\cdot;\theta)) &= \mathbb{E}_{p(z,c)}\left[\log \frac{p(z,c)}{p(z)p(c)}\right] = \\ &= \mathbb{E}_{p(z,c)}\left[\log \frac{p(z|c)}{p(z)}\right] = \\ &= \sum_{z \in \mathbb{I}, c \in \mathbb{C}} p(z,c) \cdot \left[\log \frac{p(z|c)}{p(z)}\right] = \\ &= \sum_{z \in \mathbb{I}, c \in \mathbb{C}} p(z|c) \cdot p(c) \cdot \left[\log \frac{p(z|c)}{p(z)}\right] \end{aligned} \tag{3.5}$$

The algorithm that computes the value of the mutual information in Equation (3.5) is listed in Algorithm 3.1. Note that this approach, while simple to implement, suffers from the curse of dimensionality as its complexity grows exponentially with the dimension of the information space.

The rest of the learning procedure is optimizing the value of the mutual information provided by the procedure `calcMI()` with respect to $\theta \in \Theta$. The learning procedure is described in Algorithm 3.2. Step 4 of the algorithm is carried as follows. Since the problem is non-convex, in order to optimize the objective function, we first employ the genetic algorithm with the fitness function equal to the objective function. The chosen selection policy is fitness-proportional while the chosen transformations (evolution/genetic) operators are both mutations and crossovers [Banzhaf et al., 1998]. After the genetic algorithm terminates, the toolbox engine runs a local optimization algorithm (`fmincon()`) starting from the parameters that were found by the genetic algorithm for further refinement.

Algorithm 3.2 learnPMF $(\Theta, R_{\text{gen}}, D)$ – the learning procedure

Input: Θ: parameter space.
Input: $R_{\text{gen}} : \Theta \to \mathcal{F}$: generator of PMFs from the parameter space.
Input: $D = \{(x_i, c_i)_i\}$: the input dataset.
Output: (θ^*, R^*): an optimal PMF.

1: $D_c \leftarrow \{c_i\}$
2: $h_c \leftarrow \text{hist}\,(D_c)$
3: $p(c) \leftarrow \frac{h_c}{|D_c|}$
4: $\theta^* \leftarrow \arg\min_{\theta\in\Theta} \text{calcMI}\,(R_{\text{gen}}\,(\theta)\,, D, p\,(c))$
5: $R^* \leftarrow R_{\text{gen}}\,(\theta^*)$
6: **return** (θ^*, R^*)

3.4.5 Experimentation

We now turn to to demonstrate the performance of PDI and the presented MATLAB toolbox, on a real-world dataset. for this purpose, we use a subset of the data collected in the National Health and Nutrition Examination Survey of 2012, carried out and published by the CDC.[11] Specifically, we utilize the Body Mass Measures subset of the survey.[12]

Setting

The Body Mass Measures dataset includes data pertaining to BMI and weight of individuals of both genders that are 19 years of age or younger. Therefore, this dataset is consistent with the illustrating example presented in Section 3.4.2. To recap the definitions:

- The information space, $\mathbb{I} = \{(bmi, w) \in \mathbb{R}^2\}$, consists of pairs of BMI and weight, respectively.
- The set of diagnoses is $\mathbb{C} = \{UW, HW, OW, OB\}$ for *i*) underweight; *ii*) healthy weight; *iii*) overweight; and *iv*) obese, respectively (Table 3.1).
- The dataset $D = \left\{(x_i, c_i)_{i=1}^N | x_i \in \mathbb{I}, c_i \in \mathbb{C}\right\}$ includes $N = 3355$ data points.

The data distribution, per weight category $c \in \mathbb{C}$, is depicted in Figure 3.9. Since the classification of each diagnosis depends on the age and gender of the subject, and since the subject doesn't send her or his age and gender as part of the disclosed information ($\mathbb{I}$ is limited to BMI and weight), Eve can't perfectly infer the weight category c of a subject simply from the disclosed message x. That said, Eve can still learn some information about the diagnosis if she gets to observe the original message sent by the subject, x, as shown in the next section.

Inference Based on Original Data

To demonstrate Eve's ability to perform inference, based on the raw (not sanitized) data, we trained 3 Support Vector Machine (SVM) classifiers with Gaussian (Radial Basis Function) kernels.[13] The first classifier considers $\{UW\}$ to be the "positive" class, and the rest of the weight categories to

[11] National Health and Nutrition Examination Survey: `https://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx`.

[12] Body Mass Measures: `https://wwwn.cdc.gov/nchs/nhanes/2011-2012/BMX_G.htm`.

[13] The Radial Basis Function kernel is of the form $K(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$.

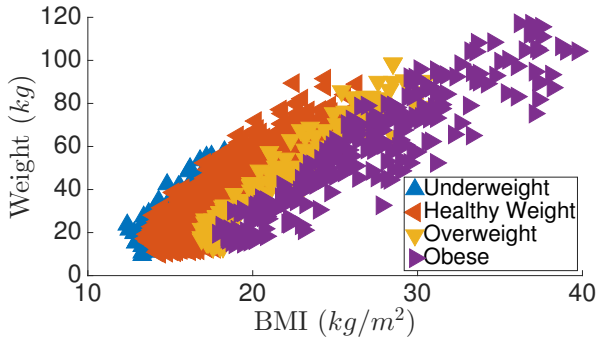


Figure 3.9: The distribution of the raw (not sanitized) data, $p(x|c)$, per weight category, $c \in \mathbb{C}$. Note that the weight categories are not perfectly separable in $\mathbb{I}$.

		True Category			
		UW	**HW**	**OW**	**OB**
Predicted Category	**UW**	47	20	0	0
	HW	14	1203	66	1
	OW	0	45	194	47
	OB	0	2	37	308

Table 3.2: The confusion matrix of the classification of the raw (not sanitized) data. UW = Underweight, HW = Healthy Weight, OW = Overweight, OB = Obese.

be the "negative" class. The second classifier, considers $\{UW, HW\}$ to be the "positive" class, and the rest of the weight categories to be the "negative" class. The last classifier, considers $\{UW, HW, OW\}$ to be the "positive" class, and $\{OB\}$ to be the "negative" class. Given a datapoint $x \in \mathbb{I}$, we predict the weight category by taking a majority vote of these 3 classifiers.

We randomly divided the dataset, by a 40 : 60 split, for training : testing, respectively. The resultant training subset included 1371 datapoints, and the resultant testing subset included 1984 datapoints. Based on the training subset, 10-fold cross-validation was utilized to optimized for the parameters of the classifiers, including the choice of each kernel's free parameter, σ. As mentioned earlier, the classification phase is performed by taking a majority vote from the 3 classifiers. The confusion matrix of the classification (on the test set) is presented in Table 3.2. As a quick measure of accuracy, the total accuracy of the classification is 88.31%.[14]

This measure indicates that about 88% of the time, Eve will guess the weight category of a subject correctly, based on a single message $x \in \mathbb{I}$ that is not sanitized. If Eve gets to see more messages from the same subject, she may be able to do even better than than this. Therefore, it is advisable to defend against this inference threat. Next section, we employ PDI to achieve that, and use the accuracy presented in this section as a benchmark to assess the performance of PDI in limiting this threat.

Inference Based on Sanitized Data

In order to sanitize the BMI and weight information that is being disclosed against the inference threat we demonstrated, we utilize PDI. In order to learn the PMF, R, we use the MATLAB toolbox presented in Section 3.4.4 (using the training subset only) [Aranki and Bajcsy, 2016]. We limit the search for a PMF, R, to the following search space that is a subset of the affine encoding functions presented in Section 3.4.4.

$$\mathcal{F} \triangleq \left\{ R : \mathbb{C} \to \mathbb{I}^{\mathbb{I}} \;\middle|\; [R(c)](x) = \begin{bmatrix} a_{c,1} & 0 \\ 0 & a_{c,2} \end{bmatrix} \cdot (x - b_c), \; b_c \in \mathbb{R}^2, a_{c,i} \geq 0.1, \forall c \in \mathbb{C}, i \in 1,2 \right\}.$$

[14] The adopted total accuracy measure is $trace(M)/N$ where M is the confusion matrix and N is the cardinality of the test set. This is the percentage of true classifications over the test set. Note that other classification performance parameters can be calculated directly from the confusion matrix.

Algorithm 3.3 The MATLAB code for learning the PMF from the BMX_G data using the PDI toolbox.

```
1  pdi_begin % Begin the definition of a PDI problem
2      % Declare data dimensions of the data
3      pdi_dimension BMI 0:2:60; % BMI is discretized by 2kg/m^2
4      pdi_dimension weight 0:5:180; % Weight is discretized by 5kg
5
6      % Declare diagnoses
7      pdi_class UW HW OW OB % Underweight; Healthy Weight; Overweight; Obese
8      % Provide data for the different diagnoses
9      pdi_datapoints UW UW_DATA
10     pdi_datapoints HW HW_DATA
11     pdi_datapoints OW OW_DATA
12     pdi_datapoints OB OB_DATA
13
14     % Declare parameters for PMF (affine transformations)
15     pdi_var shift(pdi_nrdimensions, pdi_nrclasses); % Shift parameters
16     pdi_var scale(pdi_nrdimensions, pdi_nrclasses); % Scale parameters
17
18     % Constraints on the parameters
19     scale(:,1) == 1; % Don't scale the data from the Underweight diagnosis
20     shift(:,1) == 0; % Don't shift the data from the Underweight diagnosis
21     scale >= 0.1; % Don't scale by 0 (to ensure left inverse)
22
23     % PMF: function of the parameters and diagnosis (affine transformations)
24     pdi_reference f(x, c) bsxfun(@times, ...
25         bsxfun(@minus, x, shift(:,c)), scale(:,c));
26 pdi_end % End the definition of the PDI problem and solve
```

In order to eliminate multiple solutions that sanitize in a statistically equivalent way, we also fix the encoding of the UW diagnosis to be the identity (i.e., $a_{UW,1} = a_{UW,2} = 1$ and $b_{UW} = \vec{0}$).[15] This yields the following parameter space

$$\Theta \triangleq \left\{ (a_{c,1}, a_{c,2}, b_c)_{c \in \mathbb{C} \setminus \{UW\}} | b_c \in \mathbb{R}^2, a_{c,i} \geq 0.1, \forall c \in \mathbb{C} \setminus \{UW\}, i \in 1, 2 \right\}, \tag{3.6}$$

for a total of 12 real-numbered parameters in the search space.

The MATLAB code that learns this PMF, $R \in \mathcal{F}$, is listed in Algorithm 3.3. We start by declaring a block of PDI toolbox code in line 1. The code defines the two dimensions of our information space $\mathbb{I}$ in lines $3-4$, then defines the set $\mathbb{C}$ (weight categories) in line 7. Afterwards, the code provides the data, per weight category, to the PDI toolbox in lines $9-12$ (the data, per category, are stored in the variables `UW_DATA`, `HW_DATA`, `OW_DATA` and `OB_DATA`). The code then moves to describe the parametric search space, as described in Equation (3.6). This is done by first defining the parameters of the search space in lines $15-16$, and then providing the constraints over these parameters in lines $19-21$.[16] Finally, the code ties the parameters together by defining the PMF generator function in lines $24-25$, and close the PDI toolbox code block in line 26.

[15] We regularize the sanitization of one class because if we have two PMFs, R_1 and R_2, satisfying $[R_1(c)](x) = A \cdot ([R_2(c)](x) - b)$, we will get $I(C, Z|S; R_1) = I(C, Z|S; R_2)$, yielding some degrees of freedom in our search space.

[16] Note that in an effort to make Algorithm 3.3 more uniform and easier to read, we also defined parameters for the UW weight category, and constrained them to represent the identity encoding function.

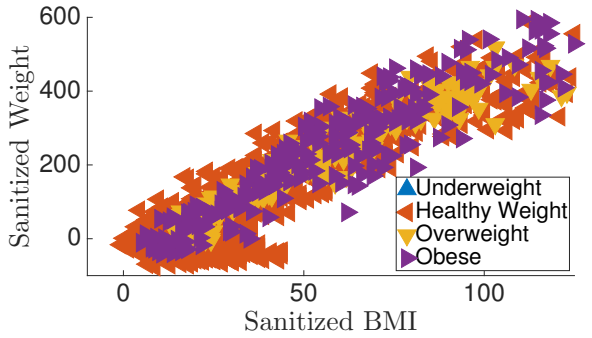


Figure 3.10: The distribution of the sanitized data, $p(z|c)$, per weight category, $c \in \mathbb{C}$. Note that the different weight categories are now less distinguishable than before.

		True Category			
		UW	**HW**	**OW**	**OB**
Predicted Category	**UW**	48	14	0	5
	HW	13	1217	276	290
	OW	0	25	13	29
	OB	0	14	0	32

Table 3.3: The confusion matrix of the classification of the sanitized data. UW = Underweight, HW = Healthy Weight, OW = Overweight, OB = Obese.

Using the resultant PMF, $R \in \mathcal{F}$, we move to sanitize the whole dataset. The distribution of the sanitized data, per weight category, is depicted in Figure 3.10. Note how the different weight categories are now less distinguishable from the sanitized data. That is, the distributions of sanitized data, per weight class category are roughly "folded," in line with our intuition from Sections 3.4.2 and 3.4.3.

In order to evaluate the goodness of the resultant sanitization, we pursue to emulate Eve by attempting to classify the weight category, using the sanitized information. For that purpose, we train 3 SVM classifiers with Gaussian (Radial Basis Function) kernels on the sanitized training subset using 10-fold cross-validation in a similar fashion to the ones described earlier. Also, similar to earlier, we use a majority vote from these classifiers to predict the weight category. The results of the classification on the sanitized test subset are presented in Table 3.3.

The total classification accuracy dropped from 88.31% to 66.03%. As would be expected, when classes are not distinguishable, it is natural to default the classification to the class with the highest number of datapoints. This intuition is verified in Table 3.3, where the classifier predicts "healthy weight" for most datapoints. Simply speaking, if a classifier would want to make a "bet" about the weight category of a subject, without any additional information, it would want to bet on the class with the highest number of subjects.

The question is, how good is this degradation from a privacy point of view, and how far is it from the best possible privacy protection? Formally, we can devise a tight lower bound on the accuracy of classifiers for our data as follows. Consider a deterministic classifier that outputs "healthy weight" for every input, regardless of what that input may be. This classifier can always be built, regardless of what sanitization process the data undergo. It's accuracy would be $1270/1984 = 64.01\%$, which is very close to our result of 66.03%. In other words, our achieved privacy protection is not very far from the best privacy protection possible.

We note that the classification results for the "underweight" category, before and after sanitization, are comparable in accuracy. This can be explained by the the choice of modeling the data distributions in the MATLAB toolbox, which is non-parametric, based on high-dimensional histograms. This type of modeling can make it hard to capture the nature of data that are rare. This is the case for the "underweight" category, where 126 datapoints (3.76%) in the dataset represent subjects from that weight category. This difficulty can be addressed by employing parametric modeling techniques for the distribution of data in different classes.

Finally, to demonstrate, in an alternative way, how this inference threat is thwarted, we take a

piece of (sanitized) information at random from our dataset, $z = [77.17, 296.45]^T$. This emulates the scenario of z being intercepted by an eavesdropper. Given that we don't know what class this datapoint came from (in order to decode it), we endeavor to try to reason about the weight category, by elimination, as follows. We will iteratively assume a weight category, c; decode the datapoint, z, under that hypothesis, and eliminate the hypothesis if the decoded datapoint $x = [R(c)]^l(z)$ is unlikely under hypothesized category c.

Healthy weight By decoding z under the assumption of $c = HW$, we retrieve $x = [21, 53.8]^T$, which is a legitimate "healthy weight" BMI and weight datapoint. Therefore, we don't eliminate the option that this datapoint came from a subject diagnosed with healthy weight.

Overweight By decoding z under the assumption of $c = OW$, we retrieve $x = [25.12, 62.4]^T$, which is also a legitimate "overweight" BMI and weight datapoint. Therefore, we don't eliminate the option that this datapoint came from a subject diagnosed with overweight.

Obese By decoding z under the assumption of $c = OB$, we retrieve $x = [30.42, 69.08]^T$, which is also a legitimate "obese" BMI and weight datapoint. Therefore, we don't eliminate the option that this datapoint came from a subject diagnosed with obesity.

In other words, Eve will not be able to rule out any c as diagnosis, since the decoded message under the assumption of diagnosis c, $x = [R(c)]^l(z)$, will be a valid datapoint from the class c, according to the generative model $p(x|c, s)$. This intuition, is the essence of how the inference threat is defended against in PDI.

3.5 Summary and Discussion

We started off by outlining the acceptability and privacy findings from two of our studies in the Berkeley Telemonitoring project: the CHF study and the RunningCoach study. We argued that the the subjects surveyed reported a level of trust towards technology researchers. That level of trust was also comparable to how much subjects trusted their physicians.

Consequently, we asked the following question in light of the responsibility that technology researchers now have when it comes to designing privacy-aware systems. What are the design principles that they technology researchers need to adopt in order to build privacy-aware systems? After reviewing the literature for existing principles and practices, we identified a new principle that is applicable to our applications in predictive medicine. Specifically, we introduced the design principle of *Inference* in an effort to encourage designers to think about the new threats to privacy as a product of machine learning and statistical inference. Simply speaking, the *Inference* principle states that the data themselves *need not* be the private object, but rather can *be used* to infer private information.

Adopting the *Inference* principle, we derived a framework that protects the individual from having their private information from being inferred from the communicated messages. We provided theoretical analysis and properties of the devised framework. We have shown, against intuition, that in our setting, one could achieve complete privacy while maintaining full utility. We stated a theorem that provides sufficient conditions for this perfect privacy-utility trade-off to occur. Moreover, we showed that such conditions are achievable by providing closed-form solutions to some cases of data generative models. It is important to observe that the perfect privacy-utility trade-off scenario is not a function of the modeling of the eavesdropper's auxiliary knowledge. This observation is important because modeling eavesdropper's auxiliary knowledge is generally a hard

problem; also, the PMF that achieves perfect privacy is the same for any eavesdropper because of the same observation. In summary, in the case of perfect privacy, the same sanitization protects patients from all eavesdroppers, regardless of their auxiliary knowledge.

Subsequently, we discussed our implementation of the learning problem resulting from the framework in a form of an open source MATLAB toolbox. We demonstrated its use with a data set published by the CDC using data about individuals' BMIs, weights and their weight categories. The experimentation shows that after sanitizing the data set, the classification accuracy drops significantly, near a lower bound of guaranteed classification accuracy, thus achieving our set goal.

Acknowledgments

Many people have directly or indirectly contributed to this chapter. First and foremost, I would like to thank my dissertation committee members, Professors Ruzena Bajcsy, John Canny and Deirdre Mulligan for their contributions, support, feedback and ideas, which greatly improved the quality of this chapter. Many thanks to go to Katherine Driggs-Campbell for the initial conversation that spurred the idea of PDI. I am also greatly indebted to Gregorij Kurillo, Yusuf Erol, Michael Carl Tschantz and Arash Nourian for the many fruitful discussions and continued feedback that improved the quality of PDI. Many of the lessons I learned about privacy came from the studies that I was part of. I am eternally grateful to David M. Liebovitz, MD, Ariane Garrett, Mita Goel, MD/MPH, Enid Montague, PhD, Robert A. Gordon, MD, Daniel Schimmel, MD, Chantal M. Mendes and the entire staff at Northwestern Medical Faculty Foundation for their tireless work that made the CHF study possible. I am also privileged to have worked alongside every single member of the amazing Berkeley Telemonitoring project team, whose quality work made the RunningCoach study possible.[17] Explicitly, I would like to thank Posu Yan, Arjun Chopra, Eugene Song, Adarsh Mani, Phillip Azar, Jochem van Gaalen, Quan Peng, Priyanka Nigam, Maya P. Reddy, Sneha Sankavaram, Qiyin Wu, Uma Balakrishnan, Hannah Sarver, Lucas Serven, Carlos Asuncion, Kaidi (Kate) Du, Caitlin Gruis, Gao Xian Peh, Yu (Sean) Xiao and Joany Gao. I would also like to thank Heather M. Patterson, Martin French and Helen Nissenbaum for their contribution in the design of the privacy and acceptability survey as well as other privacy-related aspects of the CHF study.

This work was supported in part by TRUST, Team for Research in Ubiquitous Secure Technology, which receives funding support for the National Science Foundation (NSF award number CCF-0424422). This manuscript was made possible by Grant Number HHS 90TR0003/01. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the United States Department of Health and Human Services. This work was supported in part by the Center for Long-Term Cybersecurity (CLTC) at UC Berkeley. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the CLTC.

Any errors or mistakes that made it to the final version of this chapter, including typographical ones, are solely my responsibility, not that of any person or entity mentioned above.

[17] `https://telemonitoring.berkeley.edu/team/`

Bibliography

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D. and Bajcsy, R. Private Disclosure of Information MATLAB Toolbox at `https://telemonitoring.berkeley.edu/PDI/`. 2016.

Aranki, D., Balakrishnan, U., Sarver, H., Serven, L., Asuncion, C., Du, K., Gruis, C., Peh, G. X., Xiao, Y., and Bajcsy, R. RunningCoach – Cadence Training System for Long-Distance Runners. In *Proceedings of Health-i-Coach '17*. ACM, Barcelona, Spain, May 2017.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-time tele-monitoring of patients with chronic heart-failure using a smartphone: Lessons learned. *IEEE Transactions on Affective Computing*, vol. 7(3):pp. 206–219, 2016.

Banzhaf, W., Nordin, P., Keller, R. E., and Francone, F. D. *Genetic programming: An introduction*, vol. 1. Morgan Kaufmann Publishers, Inc., 1998.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2 edn., 2006.

Cranor, L., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J., Schunter, M., Stampley, D. A., and Wenning, R. *The platform for privacy preferences 1.1 (P3P1. 1) specification*. The World Wide Web Consortium (W3C), 1.1 edn., November 2006. URL `https://www.w3.org/TR/P3P11/`.

Department of Justice's Office of Privacy and Civil Liberties (OPCL). Overview of The Privacy Act of 1974. 2015. Accessed: July 2017, URL `https://www.justice.gov/opcl/overview-privacy-act-1974-2015-edition`.

du Pin Calmon, F. and Fawaz, N. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pp. 1401–1408. IEEE, 2012.

Dwork, C. Differential privacy. In *Automata, Languages and Programming*, pp. 1–12. Springer, 2006.

European Parliament. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal L*, vol. 281:pp. 31–50, October 1995. Accessed: July 2017, URL `http://data.europa.eu/eli/dir/1995/46/oj`.

Evfimievski, A., Gehrke, J., and Srikant, R. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 211–222. ACM, 2003.

Jiao, J., Courtade, T., Venkat, K., and Weissman, T. Justification of logarithmic loss via the benefit of side information. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 946–950. IEEE, 2014.

Langheinrich, M. Privacy by design—principles of privacy-aware ubiquitous systems. In *Ubicomp 2001: Ubiquitous Computing*, pp. 273–291. Springer, 2001.

Miller, B., Huang, L., Joseph, A. D., and Tygar, J. D. I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis. In *Privacy Enhancing Technologies: 14th International Symposium, PETS 2014, Amsterdam, The Netherlands, July 16-18, 2014. Proceedings*, pp. 143–163. Springer International Publishing, 2014. ISBN 978-3-319-08506-7. doi: 10.1007/978-3-319-08506-7_8. URL `http://dx.doi.org/10.1007/978-3-319-08506-7_8`.

Narayanan, A. and Shmatikov, V. How To Break Anonymity of the Netflix Prize Dataset. *CoRR*, vol. abs/cs/0610105, 2006. URL `http://arxiv.org/abs/cs/0610105`.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE, 2008.

Nissenbaum, H. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

Pitofsky, R., Anthony, S. F., Thompson, M. W., Swindle, O., and Leary, T. B. Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress. 2000. Note: Commissioner Swindle dissented from the report and Commissioner Leary concurred in part and dissented in part.

Pitofsky, R., Azcuenaga, M. L., Anthony, S. F., Thompson, M. W., and Swindle, O. Privacy online: A report to congress. 1998.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, vol. 58(1):pp. 7–11, 2011.

Privacy Act. US Congress. *5 U.S.C.*, (§ 552a), 1974.

Rebollo-Monedero, D., Forne, J., and Domingo-Ferrer, J. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22(11):pp. 1623–1636, 2010.

Reed, I. S. Information theory and privacy in data banks. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, pp. 581–587. ACM, 1973.

Salamatian, S., Zhang, A., du Pin Calmon, F., Bhamidipati, S., Fawaz, N., Kveton, B., Oliveira, P., and Taft, N. How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data. In *GlobalSIP*, pp. 269–272. 2013.

Sankar, L., Rajagopalan, S. R., and Poor, H. V. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, vol. 8(6):pp. 838–852, 2013.

Schaar, P. Privacy by design. *Identity in the Information Society*, vol. 3(2):pp. 267–274, 2010.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10(05):pp. 557–570, 2002.

Tavani, H. T. and Bottis, M. The consent process in medical research involving DNA databanks: some ethical implications and challenges. *ACM SIGCAS Computers and Society*, vol. 40(2):pp. 11–21, 2010.

US Congress. Federal Trade Commission Act. *15 U.S.C.*, (§§ 41-58), 1914.

van Rossum, H. *Privacy enhancing technologies: the path to anonymity*. Registratiekamer, 1995.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60(309):pp. 63–69, 1965.

Warren, S. D. and Brandeis, L. D. The right to privacy. *Harvard Law Review*, pp. 193–220, 1890.

Westin, A. F. *Privacy and Freedom*. Atheneum, New York, 5 edn., 1968.

Westin, A. F. and The Staff of The Center for Social & Legal Research. Bibliography of Surveys of the U.S. Public, 1970-2003. 2003. Accessed through the Wayback Wachine (July 2017), URL `http://www.privacyexchange.org/survey/surveys/surveybibliography603.pdf`.

White, A. M., Matthews, A. R., Snow, K. Z., and Monrose, F. Phonotactic reconstruction of encrypted VoIP conversations: Hookt on fon-iks. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pp. 3–18. IEEE, 2011.

Yamamoto, H. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.). *IEEE Transactions on Information Theory*, vol. 29(6):pp. 918–923, 1983.

Chapter 4

A Data-Driven Approach to Detecting Publication Bias

> *As for the search for truth, I know from my own painful searching, with its many blind alleys, how hard it is to take a reliable step, be it ever so small, towards the understanding of that which is truly significant.*
>
> – Albert Einstein, 1934

4.1 Preface – Thought Experiment

Let us entertain the following thought experiment. We consider the following hypothesis.

Hypothesis 4.1 (H)**.** *Daniel Aranki can tell the color of any face-down playing card, merely by looking at its back.*

First, let me assure you that I, Daniel Aranki, have no psychic abilities whatsoever. In particular, I cannot tell the color of a random face-down card with a probability higher than pure luck.

The psychic community decides that they're willing to empirically accept hypothesis H, if a study protocol is devised, and carried out, such that it can distinguish the truthfulness of this hypothesis, up to a "benchmark" probability of Type I error of at most 0.001.[1] This benchmark probability is often called *the significance level*.

They come up with the following test protocol.

Study Protocol 4.1. *Draw* 10 *playing cards, face down, at random, with replacement. For each drawn card, have Daniel assert its color. Declare that hypothesis H is accepted if all of Daniel's assertions about the colors are correct.*

Let us understand whether Study Protocol 4.1 satisfies the community's standard of accepting hypothesis H. In order to do so, we devise the following opposite hypothesis (the hypothesis of the skeptical).

Hypothesis 4.2 (H_0)**.** *Daniel Aranki* cannot *tell the color of any face-down playing card merely by looking at its back.*

Hypothesis H_0 is often called the *null hypothesis*, and is the logical negation of the original hypothesis H. We note that hypothesis H is accepted if and only if hypothesis H_0 is rejected. Now we ask ourselves, what is the chance that Study Protocol 4.1 yields acceptance of hypothesis H even if it is false (i.e., Type I error)? The answer to this question dictates whether the psychic community will adopt Study Protocol 4.1 as their protocol for accepting/rejecting hypothesis H. If we denote the output of Study Protocol 4.1 by O (accept H or reject H), our question translates into calculating the probability $p(O = \text{ accept } H|H_0)$, the probability of accepting H when it is the case that H_0 is the true hypothesis.

This probability is easy to calculate. If we assume the null hypothesis H_0 is the correct hypothesis, then the probability of a correct guess of a card's color by Daniel is 0.5 (since the deck has equal number of red and black cards). Moreover, because of replacement, each guess constitutes an independent event from the rest of the guesses. Formally, if G_i is the correctness of the ith

[1] Type I error is commonly referred to as the "false positive" error. That is, accepting a hypothesis when it is actually false.

guess by Daniel, we have $p(G_i = \text{ correct } | H_0) = 0.5$ for all $i \in \{1, \ldots, 10\}$. Moreover, $G_i \perp G_j | H_0$ for all $1 \leq i < j \leq 10$. The event $\{O = \text{ accept } H\}$ is equivalent to the event $\bigcap_{i=1}^{10} \{G_i = \text{ correct}\}$. Therefore, we conclude that $p(O = \text{ accept } H | H_0) = p(G_1 = \text{ correct}, \ldots, G_{10} = \text{ correct } | H_0) = \prod_{i=1}^{10} p(G_i = \text{ correct } | H_0) = 0.5^{10} = 1/1024$.

Since this probability is smaller than the benchmark probability of 0.001 (significance level), the community adopts Study Protocol 4.1 as the protocol for accepting/rejecting hypothesis H. Even though everything may seem to be in order, according to the scientific method, we identify an issue that has serious implications on the reliability of empirical findings in science and their interpretations.

4.1.1 Selection Bias

The issue of selection bias can be explained using the following series of scenarios. Say now that different groups of people from the community implement Study Protocol 4.1, and Daniel happily obliges. You are now trying to assess for yourself, whether you want to accept hypothesis H from their reports about their trials (you are not necessarily running Study Protocol 4.1 yourself, you are now simulating a curious reader). You read the psychic literature exhaustively to count the number of successful trials of Study Protocol 4.1 (a successful run means that Daniel succeeded to guess the colors of all 10 cards in the trial). You find that 3 groups of people succeeded in their trials of Study Protocol 4.1. Would you accept H, if you wanted to adopt the benchmark probability of Type I error of at most 0.001 (significance level)? We argue that you did not gather sufficient information to make that determination. Let us see why.

Assuming the null hypothesis H_0, each run of Study Protocol 4.1 is a Bernoulli experiment with success probability of $p = 1/1024$. Moreover, under the null hypothesis H_0, different runs of Study Protocol 4.1 are independent. Therefore, in order to determine the probability of Type I error, incorrectly accepting H (equivalently, incorrectly rejecting H_0), you rely on the probability of a Bernoulli trial. That is, assuming the null hypothesis H_0, the probability of k successful experiments of Study Protocol 4.1 out of a total of n trials can be calculated by $p(k \text{ out of } n | H_0) = \binom{n}{k} p^k (1-p)^{(n-k)}$, where $\binom{n}{k}$ is the Bernoulli coefficient (n choose k).

If you are adopting the benchmark probability of Type I error of at most 0.001 (significance level), you will reject the null hypothesis, H_0, and accept the original hypothesis, H, if $p(k \text{ out of } n | H_0) \leq 0.001$. In the current setting, you only know that $k = 3$. In order to evaluate this probability, you need to either know i) the total number of trials, n; or ii) the number of unsuccessful trials, $n - k$. Therefore, you need to go back to the psychic literature and count the total number of trials (or the number of unsuccessful trials), as well.

After going back to the literature, you find that there is a total of 100 reported studies (3 of which were successful). You now calculate $p(3 \text{ out of } 100 | H_0) = \binom{100}{3} p^3 (1-p)^{97} \cong 0.00014 \leq 0.001$ and determine that you want to accept hypothesis H. Alright, now everything must be dandy, no? Unfortunately, there is no clear yes/no answer just yet. The answer to this question depends on the answer to the another, more fundamental, question.

Question 4.1. *Was every trial, successful or unsuccessful, reported in the psychic literature?*

If the answer to this question is *yes*, then your conclusion is consistent with your significance level. On the other hand, if the answer is *no*, then the conclusion may not be valid, depending on a few factors. To demonstrate this, let's ask ourselves, what would be our determination as to the acceptance of hypothesis H (equivalently, rejection of H_0) if we had an oracle that can tell us exactly how many trials were there (including unreported ones) and how many of them

True number of trials			
Total (n)	Successful (k)	Type I error	Reject H_0?
100	3	$\cong 0.00014$	Yes
200	3	$\cong 0.00101$	**No**
400	6	$\cong 3.2 \times 10^{-6}$	**Yes**

Table 4.1: Different scenarios of true number of trials and true number of successful ones, with the Type I error in each scenario and whether we reject H_0 with significance level 0.001.

were successful (including unreported ones)? We summarize the answer to this question under different scenarios in Table 4.1. The first row in the table describes the scenario where all trials were reported and you discovered them all (when the answer to Question 4.1 is yes).

Clearly, our determination depends on the true number of trials and the true number of successful trials. For example, the second row in Table 4.1 describes a scenario, in which there are 200 trials total, with only the 3 you discovered being successful and the rest being unsuccessful. In this scenario, our determination would actually be reversed and we would not accept hypothesis H (equivalently, not reject H_0). This is because $p(3 \text{ out of } 200|H_0) = \binom{200}{3}p^3(1-p)^{197} \cong 0.00101 > 0.001$. Therefore, if some trials go unreported, our determination based only on reported trials may be wrong.

The third row in Table 4.1 demonstrates a counter-example to a common, and dangerous, misconception in this realm. The misconception is that if reports from both categories (successful and unsuccessful) are censored in equal proportions, then unbiased conclusions can be drawn from the censored set. Note that the number of true total trials and true successful trials in the third row are double of the corresponding numbers in the second row (which implies that the true number of unsuccessful trials is also doubled). However, the conclusion drawn from each scenario is different. In simple words, even though the difference between the second and third rows simulates "proportional" censorship (successful and unsuccessful trials were each censored by the same factor of 50%), the resulting conclusions are opposites (and the probabilities of Type I error are orders of magnitude apart).

Question 4.1 is a dialed-down version of the question of selection bias in general, and publication bias in the case of publishing scientific results. In this chapter, we deal with the issue of *detecting* when such bias exists in a set of published empirical results. We devise a statistical test that can quantify the probability of observing a dataset of publications "at least as extreme as the one in hand" under the hypothesis of *no bias*. The lower this probability is, the less confident we are that the given dataset is a result of unbiased publications, and the more we need to be cautious when drawing conclusions from it.

4.2 Introduction

We demonstrated, albeit with a rather comical example, the implications of publication bias on the correctness of the conclusions drawn from published empirical studies. As mentioned earlier, in this chapter we examine the problem of *publication bias* in the reporting of scientific empirical studies. In particular, we tackle the question of *detecting* publication bias in a set of publications that utilize the Student t-test [Student, 1908]. Given a dataset of such publications, we *quantify the likelihood* that this dataset was generated through an unbiased process of publishing. This likelihood, in turn, can serve as a measure for us to understand the reliability of any conclusion

we may draw from these publications. But first, let us examine why this problem is relevant to predictive medicine.

This chapter, although self contained, assumes some basic knowledge in probability theory and statistical hypothesis testing (including statistical significance analysis). Some basic introduction of the premises of statistical significance analysis is presented in Section 4.2.2.

4.2.1 Predictive Medicine

So far, we have identified some of the necessary building blocks of predictive medicine. In particular, we argued that in order to realize the predictive healthcare model, we need to devise a technology for reliable health-related data collection, which can streamline the costly epidemiological studies. This lead us to the notion of health telemonitoring (Chapter 2). Then, we endeavored to understand the privacy implications and requirements of technologies pertaining to predictive medicine, particularly telemonitoring (Chapter 3).

In essence, the premise of devising such privacy-preserving technologies is to enable the discovery of risk factors of diseases and means for their prevention. In short, risk factors can be combined in predictive models that aim to estimate the risk of clinical deterioration, while the means for prevention can be used to devise effective medical intervention protocols. These interventions may be applied once such a risk is deemed to be too high. All of this is contingent on the reliability of our scientific findings, a feat that is often wrongfully and dangerously taken for granted. Ensuring the validity of scientific findings requires careful thought, study, planning, design *and uncensored discourse*.

Publication bias is one of the hurdles that threatens our ability to identify such risk factors and prevention means. This is because such bias distorts the image of science in a way that makes this it nonrepresentative of the truth. Therefore, it is vital to correct such misgiving if it exists; and the first step in solving a problem is identifying and acknowledging it. As a result, we attempt to first identify the problem of publication bias. To be clear, there is a wide acknowledgment in the scientific community that this problem exists; however, some gaps exist in utilizing formal and objective means to unequivocally verify its existence, once and for all. An analogy to this problem, without reading too much into it, is the question in computational theory of whether the two computational classes P and NP are equal.[2] Even though there is a strong belief that the answer to this question is that $P \neq NP$, there is no formal mathematical answer to this question yet.

Because of the importance of the publication bias problem and its impact on science in general, not only the branch of predictive medicine, we elect to treat this problem in its generality. Therefore, the rest of this chapter is written in that language. The reader is encouraged to draw analogies applicable to predictive medicine from the discussion.

4.2.2 Significance Analysis – Setting

A scientific hypothesis (hereafter: hypothesis) is a *verifiable* claim that may be used to explain or predict a certain phenomenon. One method of verifying hypotheses is by conducting experiments and obtaining data. The field of statistical hypothesis testing deals with the question of verifying

[2] P (deterministic, polynomial time) is the complexity class of decision problems that are solvable in *polynomial* time (in the size of the input) by a *deterministic* Turing machine. NP (non-deterministic, polynomial time) is the complexity class of decision problems that are solvable in *polynomial* time (in the size of the input) by a *non-deterministic* Turing machine.

a given hypothesis, through statistical methods. More concretely, given a hypothesis H, statistical hypothesis testing deals with the question of quantifying the likelihood of obtaining a dataset "at least as extreme as the obtained one in a given experiment," given that the hypothesis H is *false*. Before we explain why we are assuming, in the analysis, that H is false, let us explain what we mean by "at least as extreme." In the statistical hypothesis testing method, we measure a *test statistic* (call it t) from a given dataset (the measured *effect size* in the experiment). This test statistic t is itself a result of a random process, and therefore is susceptible to randomness, because of many factors, including but not limited to sampling error (because one wouldn't generally be running the experiment on the whole target population). If we understand the distribution of the test statistic (or bounds on its distribution) under the assumption that H is false, we can reason about the likelihood of observing *any* dataset that yields a test statistic at least as extreme as the observed one t, assuming H is false. In layman terms, this quantifies the likelihood of observing an effect size, in any experiment, that is at least as large as the measured one in the current experiment, even though no actual effect exists (Type I error).

Let us now explain why we assume, in the analysis, that H is false. One strong argument for this is a somewhat philosophical one. Before proving that H is true, we can only assume that the world is unchanged and that H is not necessarily true. More formally, we attempt to prove H by a statistical analogy of the "proof by contradiction." In the traditional sense, you'd assume that H is not true, and then reach a logical contradiction. Then, relying on the soundness of your underlying theory (logic, for example), you'd have to arrive at the conclusion that H has to be true. This is because, your proof essentially showed, through logical tautologies, that $(\neg H \implies (true \iff false))$ is a true statement; and if your underlying theory is sound, $(true \iff false)$ has to be a false statement (otherwise, you'd have a paradox, and your underlying theory would not be sound after all). The statement $(\neg H \implies false)$ can only be true if $(\neg H)$ is false, which is if and only if H is true.

In the statistical analogy of the proof by contradiction, one would start by defining the null hypothesis H_0 to be the logical negation of H. Then, one would assume that H_0 is true (which is equivalent to assuming H is false). Then, if one arrives at a result that is very improbable (usually, with a probability *lower* than some predefined *significance level* α), one can draw the conclusion that, with statistical significance α, H_0 can be rejected and therefore H can be accepted. Results rejecting the null hypothesis H_0 are often called *statistically significant* or simply *significant*.

With this in mind, the *general* process of statistical hypothesis testing is as follows.[3]

1. Fix a hypothesis H.
2. Define the null hypothesis H_0, the logical negation of H.
3. Select a significance level, α, that defines the (upper) threshold of rejecting H_0 by mistake (Type I error). A few popular values, for reference, are $0.05, 0.01$ and 0.001.
4. Conduct the experiment and collect data.
5. Calculate a test statistic t from the data.
6. Calculate the p-value $p = p(T \geq t|H_0)$ (or $p = p(|T| \geq |t||H_0)$, depending on the design of the experiment) through statistical significance analysis.

[3]Some statistical tests may require adjustments to the presented general process. The presented process is consistent with the method that utilizes the Student t-test.

7. If $p \leq \alpha$, reject the null hypothesis H_0, with significance level α.
8. Otherwise, draw *no conclusion*.
9. Either way, report your results by reporting at least your number of subjects and the value of your test statistic.

4.2.3 Publication Bias in General

Publication bias against publications of experiments resulting in non-significant results and misleading representation of Type I error in published studies can distort the perceptions of both the scientific world and the public. The importance of the issue of publication bias has been identified almost four decades ago. Rosenthal famously described "the extreme view of the 'file drawer problem' is that journals are filled with the 5% of the studies that show Type I errors, while the drawers are filled with the 95% of the studies that show non-significant results" [Rosenthal, 1979].

The ability to replicate and reproduce experimental results is arguably the most important tenet of the experimental science. In a 2013 article, Nature News reported that the National Institutes of Health (NIH) was contemplating changes to their grant applications that would require applicants to validate some experimental procedures and results "in certain types of sciences, such as the foundational work that leads to costly clinical trials" [Wadman, 2013]. NIH convened two workshops in 2012 to examine the issue of reproducibility. Furthermore, the leaders of NIH, and others, published a call for higher standards in reporting preclinical research in order to optimize their predictive value. Even though their call primarily targeted the animal-related research community, they noted that the life sciences community in general "often lack[s] adequate reporting on the design, conduct and analysis of the experiments" [Landis et al., 2012]. Of the many dangers of lack of reproducibility in experimental research is that "some non-reproducible preclinical papers had spawned an entire field with hundreds of secondary publications that expanded on elements of the original observation." Even more seriously, "some of the research has triggered a series of clinical studies – suggesting that many patients had subjected themselves to a trial of a regimen or agent that probably wouldn't work" [Begley and Ellis, 2012].

More recently, the American Statisticians Association published a statement outlining the principles governing the use and interpretation of p-values [Wasserstein and Lazar, 2016; Baker, 2016]. The statement came as a result of the association's concern about issues of *reproducibility* and *replicability* of scientific conclusions, and the misunderstanding and misuse of statistical inference as a cause of the "reproducibility crisis" [Peng, 2015].

Confirming and estimating publication bias is a non-trivial task. Many of the studies conducted to confirm publication bias were based on surveying researchers. For example, in an effort to estimate publication bias–either by choosing not to submit or getting the submission rejected for publication, Dickersin et al. surveyed 318 authors to inquire whether they had participated in any unpublished randomized clinical trials. They concluded that the major reasons for non-publication were "negative" results and lack of interest [Dickersin et al., 1987].

Some others attempt to confirm (and estimate) publication bias by estimating the extent of bias and/or correct for it. In particular, many methods propose ways to test for the significance of publication bias based on their estimates of selection, provided that the underlying assumptions of their methods are satisfied. Copas argues that "correcting for this bias is not possible without making untestable assumptions" [Copas, 1999]. Therefore, methods for *detecting* publication bias that are based on *estimating* or *correcting for* publication bias will suffer from the curse of unverifiable assumptions.

In contrast, in this chapter we tackle the problem of directly detecting publication bias in a formal, data-driven way without the need to estimate such bias. We devise a statistical test aimed at detecting publication bias in a set of observed publications reporting p-values that are results of applying the Student t-test [Student, 1908]. We demonstrate the method by applying the test to a set of 3,721 publications in the field of experimental psychology, collected from nineteen journals from the American Psychological Association (APA), published between the years 2002 and 2012.

4.2.4 Chapter Organization and Contributions

The rest of this chapter is organized as follows. In Section 4.3, we review the literature for related work in the field. We then present, in Section 4.4, our formal method for quantifying the probability of observing a dataset of publications of Student t-test that is "at least as extreme as the one in hand," assuming an unbiased publication process. This quantification can, in turn, serve as a detection tool for publication bias in a set of publications. We follow this by presenting the implementation details of this method through computational means in Section 4.5. In Section 4.6, we apply our method to a large dataset of publications from the APA journals and follow this by a discussion of our results in Section 4.7.

The contributions of this chapter are i) the presentation of a formal data-driven method for the detection of publication bias; and ii) the presentation of a MATLAB toolbox that implements the aforementioned method.

4.3 Related Work

As mentioned earlier, many methods attempt to verify or estimate publication bias by surveying authors and researchers. For example, Easterbrook et al. conducted a study in which they contacted and surveyed (by interview or a questionnaire) 216 principal investigators of 487 studies. According to their findings, the authors "confirm a systematic selection bias in the publication process according to study results. Studies with a statistically significant result for the main outcome of interest were more likely to be submitted for publication and more likely to be published than studies with null results" [Easterbrook et al., 1991]. Even more research has been conducted with that emphasis. A survey of literature concerning these studies include Smith [1980], Coursol and Wagner [1986], Dickersin et al. [1987] and others.

In the area of estimating the extent of publication bias, or estimating the true effect size by accounting for publication bias, many methods were proposed. The trim and fill method is widely used in adjusting for publication bias [Egger et al., 1997; Sutton et al., 2000; Terrin et al., 2003]. This methods is appealing for its simplicity to non-statisticians, and is based on a popular graphical tool called the funnel plot. The trim and fill method makes a strong unverifiable assumption of symmetry [Egger et al., 1997; Copas, 1999; Sutton et al., 2000].

Other techniques for estimating publication bias and/or the true effect size are based on maximum likelihood estimates of selection (or selectivity) models for publication, based on the significance of the reported results [Hedges, 1992; Iyengar and Zhao, 1994; Hedges and Vevea, 1996; Copas, 1999; Ioannidis and Trikalinos, 2007]. These methods impose modeling assumptions on the distribution of the underlying effect sizes [Hedges, 1992; Iyengar and Zhao, 1994; Copas, 1999], homogeneity of studies in question [Ioannidis and Trikalinos, 2007], the underlying publication selection model [Iyengar and Zhao, 1994; Copas, 1999; Ioannidis and Trikalinos, 2007], the set of possible outcomes of each study [Ioannidis and Trikalinos, 2007] and others.

Selection models take publication bias into account by weighing the different end points based on the p-value obtained by the statistical test run on them. Hedges argued that, based on the psychological research on the interpretation of the results of statistical analyses [Rosenthal and Gaito, 1963, 1964; Nelson et al., 1986, for example], "researchers' perceptions about the conclusiveness of research results is [sic] strongly related to the p-value" [Hedges, 1992]. Hedges further argued that "a third finding is that the relationship between perceived conclusiveness of studies and p-values is not smooth but is subject to 'cliff effects' near conventionally used a priori levels of significance such as $\alpha = 0.05$ and $\alpha = 0.01$" [Hedges, 1992].

Based on these observations, Hedges introduced a selection model as a step function of the p-value, with potentially multiple discontinuities (steps) at points determined a priori. This weight function has its value between each two consecutive steps constant, and these constants are learned in a maximum likelihood manner, based on a probabilistic model of the true effect of studies (mainly studies concerned with the same underlying condition) [Hedges, 1992]. Numerical and convergence issues arise by using this, and similar, methods [Terrin et al., 2003].

Begley and Ellis published an article calling to raise the standards. The article included the findings of a study on the reproducibility of 53 preclinical cancer studies that were "deemed landmark studies". The papers "were deliberately selected that described something completely new, such as fresh approaches to targeting cancers or alternative clinical uses for existing therapeutics." They found that only 6 of them (11%) were reproducible [Begley and Ellis, 2012]. The definition of reproducibility, in their report, is based on whether the findings are sufficiently robust to drive a drug-development program.

In a 2011 study, researchers from Bayer HealthCare reported on their trials to reproduce 67 projects. Their findings are consistent with those of Begley and Ellis [2012]. The authors report that only 14 of the 67 findings in question were completely reproducible, 43 were inconsistent with their in-house experiments, 5 had main datasets that were reproducible, 3 were partially reproducible and 2 that "were almost exclusively based on in-house data." The authors estimate the complete reproducibility of the findings included in their study to be $\sim 20 - 25\%$ [Prinz et al., 2011]. Of the methods used to reproduce results, the original models were exactly copied in 12 cases, in 38 cases the models were adapted to internal needs, the "published data was transferred to models for other indications" in 2 cases, and in 5 other cases the general hypotheses of the projects could not be verified.

Pocock et al. surveyed 45 reports of comparative trials published in the *British Medical Journal*, *Lancet* and the *New England Journal of Medicine*. The survey focused on studying some of the procedures in conducting and designing clinical trials and reporting their results, such as, multiple end points, subgroup analysis, repeated measurements over time, multiple treatment groups, the excessive use of significance testing, the lack of reporting measured effect sizes (test statistics) and confidence intervals, the choice of the number of subjects and rules of early termination of trials, the potential selectivity over which significance tests to even report and the biased selection of results for the summary [Pocock et al., 1987].

4.4 Formal Method

Setting We treat hypotheses tests that utilize the Student t-test [Student, 1908]. In our setting, we assume that a set of M hypotheses tests $D = (T_i, n_i, P_i)_{i=1}^{M}$ is observed. For each test $i \in \{1, \ldots, M\}$, there exists a true (but hidden) effect size e_i. Given that effect size and the Degrees of Freedom (DOF) in the experiment, n_i, the measured t statistic of the experiment, T_i, is distributed

according to the non-central Student t distribution $T_i|n_i, e_i \sim NCT(n_i, e_i)$ [Johnson and Welch, 1940]. Moreover, given the effect sizes and DOFs of any two experiments, i and j ($i \neq j$), T_i and T_j become independent. The reported p-value, P_i, of the experiment i follows the following relationship

$$P_i = 2P(T \geq T_i; n_i) \tag{4.1}$$

where T is distributed according to the Student t distribution (with n_i DOF). Note that the p-value as defined here ranges from 0 to 2. We elect to use this definition because it allows us to differentiate between positive and negative effect sizes directly from the p-values domain.

Null hypothesis Fix the values $0 \leq \bar{p}_1 < \bar{p}_2 < \bar{p}_3 < \bar{p}_4 < \bar{p}_5 \leq 2$. We formalize the hypothesis, $H_0\left(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5\right)$ (in short H_0) as the absence of bias against publications with p-values in the range $[\bar{p}_3, \bar{p}_5)$ as opposed to the range $[\bar{p}_1, \bar{p}_3)$ with emphasis on relative amounts of publication of p-values in the range $[\bar{p}_3, \bar{p}_4)$ to $[\bar{p}_2, \bar{p}_3)$.

Note that fixing the null hypothesis H_0 needs to be done before performing the test or the analysis, for validity. As such, fixing the values of $\bar{p}_j, j = 1, \ldots, 5$ needs to be done prior to the rest of the analysis, including exploring the data to be tested. One may elect to extract a random subsample of the dataset for exploration purposes, which should not be included in the actual test.

Test statistic design We start with some intuition regarding publication bias. Consider the probability density function (PDF) of p-values, $f(p)$. Bias against publications with p-values in the range $[\bar{p}_3, \bar{p}_5)$ relative to publications with p-values in the range $[\bar{p}_1, \bar{p}_3)$ is manifested in a significantly smaller area $\int_{\bar{p}_3}^{\bar{p}_5} f(p)dp$ compared to $\int_{\bar{p}_1}^{\bar{p}_3} f(p)dt$ (Figure 4.1).

The goal is to design a test statistic that can measure this effect in a manner that yields a probabilistic bound on observing a dataset at least as extreme as dataset D, under the null hypothesis H_0.

We define the filter

$$\mathcal{F}(p) \triangleq -1 \cdot \mathbb{I}_{[\bar{p}_1, \bar{p}_2)}(p) + 3 \cdot \mathbb{I}_{[\bar{p}_2, \bar{p}_3)}(p) - 3 \cdot \mathbb{I}_{[\bar{p}_3, \bar{p}_4)}(p) + 1 \cdot \mathbb{I}_{[\bar{p}_4, \bar{p}_5)}(p)$$
$$\text{where } \mathbb{I}_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

For intuition, this filter is a step function approximation of the third derivative of the Gaussian filter centered at $\bar{p}_3$ (Figure 4.1). Using the filter $\mathcal{F}$, we define the test statistic

$$S \triangleq \sum_{i=1}^{M} \mathcal{F}(P_i).$$

Intuitively, the filter $\mathcal{F}$ acts as follows. Datasets with bias against publications with p-values in the range $[\bar{p}_3, \bar{p}_5)$ relative to those with p-values in the range $[\bar{p}_1, \bar{p}_3)$ are likely to yield a higher value of the test statistic S than datasets without such bias. For intuition, the test statistic, acting on the PDF $f(p)$, can be calculated as $s = M \cdot \int_{\bar{p}_1}^{\bar{p}_5} f(p) \cdot \mathcal{F}(p)dp$. PDFs that are a result of such publication bias are likely to yield a higher test statistic s than PDFs that are a result of no bias. This scenario is depicted in Figure 4.1.

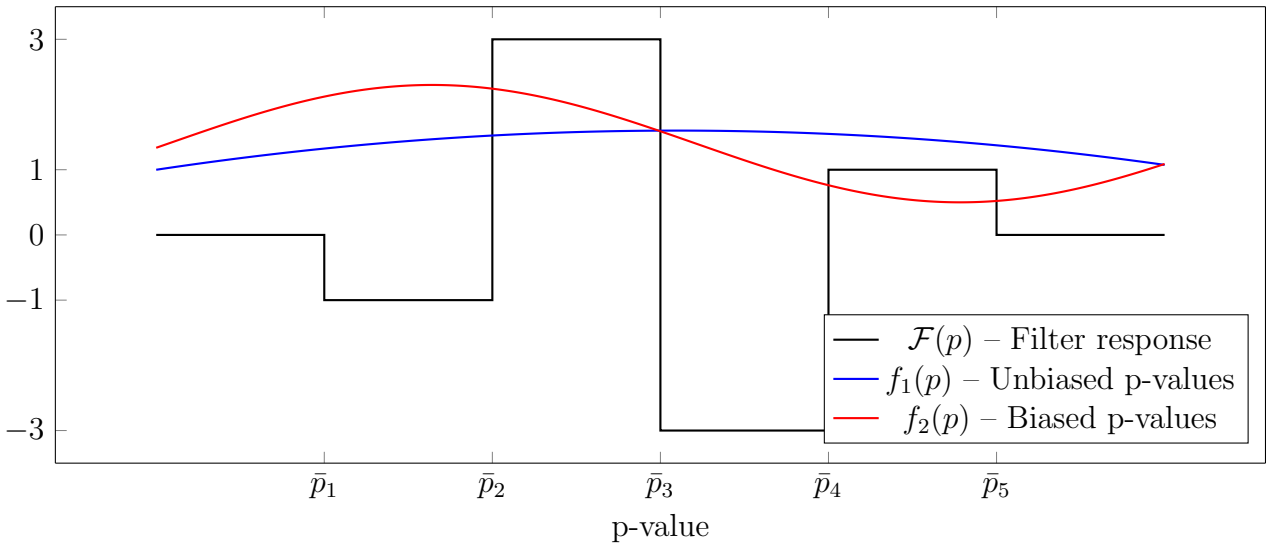


Figure 4.1: Visualization of the filter $\mathcal{F}(\cdot)$ with illustrative plots of p-values' PDFs $f_1(\cdot)$ and $f_2(\cdot)$ demonstrating cases of no bias and bias, respectively.

The statistical test We now turn to calculating an upper bound on the likelihood that we observe a dataset at least as extreme as D under the null hypothesis H_0. We calculate a bound on the tail probabilities of the random variable S. To do so, consider the random variables

$$X_i(T_i, n_i, P_i) \triangleq \begin{cases} -1 & P_i \in [\bar{p}_1, \bar{p}_2) \\ 3 & P_i \in [\bar{p}_2, \bar{p}_3) \\ -3 & P_i \in [\bar{p}_3, \bar{p}_4) \\ 1 & P_i \in [\bar{p}_4, \bar{p}_5) \\ 0 & \text{otherwise} \end{cases} = \begin{cases} -1 & T_i \in (\bar{t}_{2,i}, \bar{t}_{1,i}] \\ 3 & T_i \in (\bar{t}_{3,i}, \bar{t}_{2,i}] \\ -3 & T_i \in (\bar{t}_{4,i}, \bar{t}_{3,i}] \\ 1 & T_i \in (\bar{t}_{5,i}, \bar{t}_{4,i}] \\ 0 & \text{otherwise} \end{cases}, i \in \{1, \dots, M\} \tag{4.2}$$

where $\bar{t}_{j,i}$ is the t-score that yields a two-tailed p-value $\bar{p}_j$ given the DOF n_i. Formally, the $\bar{t}_{j,i}$ that achieves $2P(T \geq \bar{t}_{j,i}; n_i) = \bar{p}_j, j \in \{1, \dots, 5\}$. Equivalently, $\bar{t}_{j,i} = -\operatorname{tinv}(\frac{\bar{p}_j}{2}; n_i)$, where $\operatorname{tinv}(\cdot)$ is the inverse t-distribution cumulative distribution function (CDF).

From the definition in Equation (4.2) and the fact that $T_i | n_i, e_i \sim NCT(n_i, e_i)$, the probability distributions of X_i are as follows:

$$P(X_i = x | n_i, e_i) = \begin{cases} \int_{\bar{t}_{2,i}}^{\bar{t}_{1,i}} NCT(t; n_i, e_i) dt & x = -1 \\ \int_{\bar{t}_{3,i}}^{\bar{t}_{2,i}} NCT(t; n_i, e_i) dt & x = 3 \\ \int_{\bar{t}_{4,i}}^{\bar{t}_{3,i}} NCT(t; n_i, e_i) dt & x = -3 \\ \int_{\bar{t}_{5,i}}^{\bar{t}_{4,i}} NCT(t; n_i, e_i) dt & x = 1 \\ 1 - \int_{\bar{t}_{5,i}}^{\bar{t}_{1,i}} NCT(t; n_i, e_i) dt & x = 0 \end{cases} \tag{4.3}$$

Note that the test statistic random variable can be written as $S = \sum_{i=1}^{M} X_i$. We denote the probability distribution in Equation (4.3) by $\mathrm{XD}(n_i, e_i)$; as such, $X_i | n_i, e_i \sim \mathrm{XD}(n_i, e_i)$. Note that given DOFs and effect sizes, X_i and X_j are independent for any $i \neq j$.

Using the Chernoff bound, we have, for all $\theta \geq 0$,

$$p(S \geq s|H_0) \leq \frac{\mathbb{E}\exp(\theta S)}{\exp(\theta s)} = \frac{\mathbb{E}\exp\left(\theta \sum_{i=1}^{M} X_i\right)}{\exp(\theta s)} = \frac{\prod_{i=1}^{M} \mathbb{E}\exp(\theta X_i)}{\exp(\theta s)}$$

Note that the effect sizes e_i are latent and as such, it is desirable to find a global bound that is not sensitive to the true values of the e_i. This is because, otherwise, we would have to make assumptions about the distribution of e_i, which are, by large, unverifiable. In order to alleviate this difficulty, we consider the following analysis.

We denote the set of DOFs in the datapoints in the set D as N. For any given $\theta \geq 0$ and $n \in N$, we define $\tilde{X}(\theta, n) \sim \mathrm{XD}(n, \cdot)$ to be a random variable such that $\mathbb{E}\exp(\theta X_i) \leq \mathbb{E}\exp\left(\theta \tilde{X}(\theta, n)\right)$ for all $i \in \{1, \ldots, M\}$ with $n_i = n$. Formally, if we define

$$\tilde{X}(\theta, n) \triangleq \underset{X \in \{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}\}}{\arg\max} \mathbb{E}\exp(\theta X),$$

we achieve that for all n, $\mathbb{E}\exp(\theta X_i) \leq \mathbb{E}\exp\left(\theta \tilde{X}(\theta, n)\right)$ for all $i \in \{1, \ldots, M\}$ with $n_i = n$, as desired.

Furthermore, we define $M(n)$ to be the number of datapoints in D with DOF n. That is, $M(n) \triangleq \sum_{i=1}^{M} \mathbb{I}_{\{n\}}(n_i), \forall n \in N$. Now we can write,

$$p(S \geq s|H_0) \leq \frac{\prod_{i=1}^{M} \mathbb{E}\exp(\theta X_i)}{\exp(\theta s)} \leq \frac{\prod_{i=1}^{M} \mathbb{E}\exp\left(\theta \tilde{X}(\theta, n_i)\right)}{\exp(\theta s)} = \frac{\prod_{n \in N} \left(\mathbb{E}\exp\left(\theta \tilde{X}(\theta, n)\right)\right)^{M(n)}}{\exp(\theta s)}$$

This bound is valid for every $\theta \geq 0$. Therefore, we can write

$$p(S \geq s|H_0) \leq \min_{\theta \geq 0} \frac{\prod_{n \in N} \left(\mathbb{E}\exp\left(\theta \tilde{X}(\theta, n)\right)\right)^{M(n)}}{\exp(\theta s)}.$$

By plugging in our definition for $\tilde{X}(\theta, n)$, we conclude that

$$p(S \geq s|H_0) \leq \min_{\theta \geq 0} \frac{\prod_{n \in N} \max_{X_n \in \{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}\}} \left(\mathbb{E}\exp(\theta X_n)\right)^{M(n)}}{\exp(\theta s)}, \tag{4.4}$$

which gives us an upper bound on the probability that we observe any set of M publications– which includes $M(n)$ publications with DOF n for each $n \in N$, yielding a test statistic S at least as extreme as s under the null hypothesis H_0. Note that the bound in Equation (4.4) is no longer sensitive to the latent effect sizes e_i, and can be calculated only using observed variables.

4.5 Implementation

Last section, we devised a statistical test that, given a dataset $D = (T_i, n_i, P_i)_{i=1}^{M}$, yields a bound on the probability of observing any other dataset at least as extreme as D, under the null hypothesis H_0. That bound is described in Equation (4.4). In this section, we discuss an implementation of this calculation, which we have done in MATLAB. That is, we describe a MATLAB toolbox that,

given a dataset $D = (T_i, n_i, P_i)_{i=1}^{M}$, it i) calculates the test statistic s; and ii) calculates an upper bound on the probability of observing a dataset at least as extreme as D, under the null hypothesis H_0; namely, $p(S \geq s|H_0)$.

There are two details to address in implementing the bound described in Equation (4.4). The first detail is devising a method that, given a Chernoff parameter θ, calculates the value of $\mathbb{E}\exp(\theta X(n,e))$.

Second, we note that for large datasets, the inner optimization problem described as

$$\prod_{n\in N} \max_{X_n \in \{Z \sim \mathrm{XD}(n,e_\theta)|e_\theta \in \mathbb{R}\}} (\mathbb{E}\exp(\theta X_n))^{M(n)} \tag{4.5}$$

may be computationally expensive. Therefore, we address the question of an efficient implementation of it.

We will start off by describing our method for calculating the expected value $\mathbb{E}\exp(\theta X)$ in Section 4.5.1. Afterwards, we tackle the issue of efficient implementation, proving that it still yields a valid upper bound in Section 4.5.2.

4.5.1 Computing $\mathbb{E}\exp(\theta X(n,e))$

The first step is to be able to calculate the expected value

$$\mathbb{E}\exp(\theta X(n,e))$$

given the DOF n, some value for the effect size e and a value for the Chernoff free parameter θ.

We note that the random variable $X(n,e) \sim \mathrm{XD}(n,e)$, and, as such, is a discrete random variable. Therefore, the expected value of it should simply be

$$\begin{aligned}\mathbb{E}\exp(\theta X(n,e)) &= \sum_{x\in\{-1,3,-3,1,0\}} \exp(\theta x)\, p(X = x|n,e) = \\ &= \sum_{x\in\{-1,3,-3,1,0\}} \exp(\theta x)\, p_x \end{aligned} \tag{4.6}$$

where

$$p_{-1} = \int_{\bar{t}_2}^{\bar{t}_1} \mathrm{NCT}(t;n,e)dt, \qquad p_3 = \int_{\bar{t}_3}^{\bar{t}_2} \mathrm{NCT}(t;n,e)dt,$$
$$p_{-3} = \int_{\bar{t}_4}^{\bar{t}_3} \mathrm{NCT}(t;n,e)dt, \qquad p_1 = \int_{\bar{t}_5}^{\bar{t}_4} \mathrm{NCT}(t;n,e)dt$$

and

$$p_0 = 1 - \sum_{x\in\{-1,3,-3,1\}} p_x,$$

where $\bar{t}_j = -\operatorname{tinv}(\frac{\bar{p}_j}{2}; n), \forall j \in \{1,\ldots,5\}$.

The procedure implementing Equation (4.6) is listed in Algorithm 4.1.

Algorithm 4.1 calcExpExp (θ, n, e, H_0) – Calculates Equation (4.6).

Input: θ: the value of the free parameter in Chernoff bound.
Ensure: $\theta \geq 0$.
Input: n: a DOF.
Ensure: $n > 0$.
Input: e: a value for the effect size.
Ensure: $e \in \mathbb{R}$.
Input: $H_0\,(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5)$: the null hypothesis.
Ensure: $0 \leq \bar{p}_1 < \bar{p}_2 < \bar{p}_3 < \bar{p}_4 < \bar{p}_5 \leq 2$.
1: **for all** $j \in \{1, \ldots, 5\}$ **do**
2: $\quad \bar{t}_j \leftarrow -\operatorname{tinv}(\frac{\bar{p}_j}{2}; n)$
3: **end for**
4: $p_{-1} \leftarrow \int_{\bar{t}_2}^{\bar{t}_1} \mathrm{NCT}(t; n, e)dt$
5: $p_3 \leftarrow \int_{\bar{t}_3}^{\bar{t}_2} \mathrm{NCT}(t; n, e)dt$
6: $p_{-3} \leftarrow \int_{\bar{t}_4}^{\bar{t}_3} \mathrm{NCT}(t; n, e)dt$
7: $p_1 \leftarrow \int_{\bar{t}_5}^{\bar{t}_4} \mathrm{NCT}(t; n, e)dt$
8: $p_0 \leftarrow 1 - \sum_{x \in \{-1,3,-3,1\}} p_x$
9: $r \leftarrow \sum_{x \in \{-1,3,-3,1,0\}} \exp(\theta \cdot x) \cdot p_x$
10: **return** r

4.5.2 Partitioning the DOFs

We first note that in order to implement Equation (4.5) in MATLAB, as is, we either have to implement it through a loop, which impedes its convergence, or translate this into one large vectorized optimization problem as follows.

$$\max_{\substack{(X_n)_{n\in N} \text{ s.t.} \\ X_n \in \{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}\}, \forall n \in N}} \prod_{n \in N} \left(\mathbb{E} e^{\theta X_n}\right)^{M(n)}$$

which is an optimization problem with $|N|$ optimization variables (one variable for effect size e, for each X_n, $n \in N$). For large datasets, convergence of this optimization problem can be very slow, as will be demonstrated in Section 4.6.

We will develop a method that drops this complexity down, while still yielding a valid Chernoff bound, with the potential cost of its tightness. The reason we say potential, is because this method covers the original optimization problem described above as a special case; which means that the user can elect to not lose tightness in the bound, if she or he wishes.

First, let us generalize a previous definition. For any given $\theta \geq 0$ and $\hat{N} \subset N$, we define $\tilde{X}(\theta, \hat{N}) \sim \mathrm{XD}(n, \cdot)$ to be a random variable such that $\mathbb{E} \exp(\theta X_i) \leq \mathbb{E} \exp\left(\theta \tilde{X}(\theta, \hat{N})\right)$ for all $i \in \{1, \ldots, M\}$ with $n_i \in \hat{N}$. Formally, if we define

$$\tilde{X}(\theta, \hat{N}) \triangleq \underset{X \in \left\{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}, n \in \hat{N}\right\}}{\arg\max} \mathbb{E} \exp(\theta X)$$

we achieve that for any $\hat{N} \subset N$, $\mathbb{E} \exp(\theta X_i) \leq \mathbb{E} \exp\left(\theta \tilde{X}(\theta, \hat{N})\right)$ for all $i \in \{1, \ldots, M\}$ with $n_i \in \hat{N}$, as desired. Now our $\tilde{X}(\theta, \hat{N})$ is a maximizer of $\mathbb{E} \exp(\theta X)$ for all $n \in \hat{N}$, instead of just

a single DOF like before. Note that we can reproduce the old definition of $\tilde{X}$ by simply having $\hat{N} = \{n\} \subset N$ be a singleton.

Second, we also generalize $M(\cdot)$ to count the number of datapoints in D for which the DOF is in a subset, $\hat{N}$, of the degrees of freedom. Formally, we define $M(\hat{N}) \triangleq \sum_{i=1}^{M} \mathbb{I}_{\hat{N}}(n_i)$ for any $\hat{N} \subset N$. We now move to rewrite our bound Equation (4.4), but first introduce a definition from set theory.

Definition 4.1 (Set Partition)**.** *Let N be a set, and $\mathcal{N}_p = \{N_1, \ldots, N_k\}$ be a collection of sets. We call $\mathcal{N}_p$ a* partition of N *if*

1. *$N_i \neq \emptyset$ for all $i \in \{1, \ldots, k\}$;*
2. *$N_i \cap N_j = \emptyset$ for all $1 \leq i < j \leq k$; and*
3. *$\bigcup_{i=1}^{k} N_i = N$.*

Moreover, the sets $N_i, \forall i \in \{1, \ldots, k\}$ are called the partition blocks*.*

Given any partition $\mathcal{N}_p$ of the observed DOF N, we can rewrite our bound Equation (4.4) as

$$p(S \geq s|H_0) \leq \frac{\prod_{i=1}^{M} \mathbb{E} \exp(\theta X_i)}{\exp(\theta s)} \leq \frac{\prod_{i=1}^{M} \mathbb{E} \exp\left(\theta \tilde{X}(\theta, n_i)\right)}{\exp(\theta s)} = \\ = \frac{\prod_{\hat{N} \in \mathcal{N}_p} \left(\mathbb{E} \exp\left(\theta \tilde{X}(\theta, \hat{N})\right)\right)^{M(\hat{N})}}{\exp(\theta s)}.$$

Just like before, since this bound is valid for all $\theta \geq 0$, we can write

$$p(S \geq s|H_0) \leq \min_{\theta \geq 0} \frac{\prod_{\hat{N} \in \mathcal{N}_p} \left(\mathbb{E} \exp\left(\theta \tilde{X}(\theta, \hat{N})\right)\right)^{M(\hat{N})}}{\exp(\theta s)}.$$

By plugging in our definition for $\tilde{X}(\theta, \hat{N})$, we conclude that

$$p(S \geq s|H_0) \leq \min_{\theta \geq 0} \frac{\prod_{\hat{N} \in \mathcal{N}_p} \max_{\tilde{X} \in \left\{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}, n \in \hat{N}\right\}} \left(\mathbb{E} \exp\left(\theta \tilde{X}\right)\right)^{M(\hat{N})}}{\exp(\theta s)}. \tag{4.7}$$

The inner product in Equation (4.7) is now

$$\prod_{\hat{N} \in \mathcal{N}_p} \max_{\tilde{X} \in \left\{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}, n \in \hat{N}\right\}} \left(\mathbb{E} \exp\left(\theta \tilde{X}\right)\right)^{M(\hat{N})},$$

which can be vectorized, just like before, as

$$\max_{\substack{(X_{\hat{N}})_{\hat{N} \in \mathcal{N}_p} \text{ s.t.} \\ X_{\hat{N}} \in \left\{Z \sim \mathrm{XD}(n,e) | e \in \mathbb{R}, n \in \hat{N}\right\}, \forall \hat{N} \in \mathcal{N}_p}} \prod_{\hat{N} \in \mathcal{N}_p} \left(\mathbb{E} \exp\left(\theta X_{\hat{N}}\right)\right)^{M(\hat{N})}, \tag{4.8}$$

Algorithm 4.2 calcProdMaxExpExp $(\theta, D, H_0, \mathcal{N}_p)$ – Calculates Equation (4.8) using a partition of N.

Input: θ: the value of the free parameter in Chernoff bound.
Ensure: $\theta \geq 0$.
Input: $D = (T_i, n_i, P_i)_{i=1}^M$: the dataset of the Student t-test.
Input: $H_0\left(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5\right)$: the null hypothesis.
Ensure: $0 \leq \bar{p}_1 < \bar{p}_2 < \bar{p}_3 < \bar{p}_4 < \bar{p}_5 \leq 2$.
Input: $\mathcal{N}_p$: a collection of sets.
Ensure: $\mathcal{N}_p$ a partition of the set of degrees of freedom $N = \bigcup_{i=1}^M \{n_i\}$.

```
r ← 1
for all N̂ ∈ N_p do
    c ← Σ_{i=1}^M 𝕀_N̂ (n_i)
    p ← max_{e∈ℝ, n∈N̂} calcExpExp(θ, n, e, H_0)
    r ← r · p^c
end for
return r
```

and has at most $2|\mathcal{N}_p|$ variables (one e effect size per partition block and one n DOF per partition block that is not a singleton). This can greatly reduce the total number of variables in the optimization problem from $|N|$, which can be in the order of 10^3 to as low as 2 optimization variables (if the partition is $\mathcal{N}_p = \{N\}$), or almost anything in between.

The procedure that calculates the vectorized optimization problem described in Equation (4.8) is listed in Algorithm 4.2. For legibility, we elected to write the procedure in this document using a loop, instead of vectorizing it, but we note that the procedure in the MATLAB toolbox is actually vectorized.

Obviously, the resulting bound in Equation (4.7) is still a valid Chernoff bound, however it may be less tight than the original bound in Equation (4.4), depending on the choice of the partition, $\mathcal{N}_p$. For example, we note that by picking the partition $\mathcal{N}_p = \{\{n\} | n \in N\}$, the partition of singleton blocks, the bound in Equation (4.7) reduces back to the bound in in Equation (4.4), so that there is no loss of generality. Finally, the full procedure implementing the general bound in Equation (4.7) is listed in Algorithm 4.3.

Algorithm 4.3 ttestDatasetBound $(D, H_0, \mathcal{N}_p)$ – Calculates an upper bound of $p(D|H_0)$

Input: $D = (T_i, n_i, P_i)_{i=1}^M$: the dataset of the Student t-test.
Input: $H_0\left(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5\right)$: the null hypothesis.
Ensure: $0 \leq \bar{p}_1 < \bar{p}_2 < \bar{p}_3 < \bar{p}_4 < \bar{p}_5 \leq 2$.
Input: $\mathcal{N}_p$: a collection of sets.
Ensure: $\mathcal{N}_p$ a partition of the set of degrees of freedom $N = \bigcup_{i=1}^M \{n_i\}$.

```
s ← Σ_{i=1}^M X_i (calculate the test statistic)
p ← min_{θ≥0} calcProdMaxExpExp(θ, D, H_0, N_p) / exp(θ·s)
return p
```

4.6 Experiment

Significance level Before we start with any analysis, we set our significance level $\alpha = 0.01$ (1 in 100). As mentioned in Section 4.2.2, this sets the probability of Type I error at 0.01. We reject the null hypothesis H_0 (which states that there is no publication bias) whenever the bound is less than α. Therefore, the probability that we falsely reject H_0 (i.e., falsely declare publication bias) is at most 0.01.

Data collection The dataset used in this analysis was collected by automatically crawling all publications of the nineteen journals from the APA between the years 2002 and 2012.[4] Consistent with the method, we only considered publications reporting p-values that were resulted of using the Student t-test [Student, 1908]. We extracted, for each reported Student t-test, the t-score, the degrees of freedom and the reported p-value.

The APA publication manual has a strict format of reporting a Student t-test resulting p-value in a publication. The format is $t(\mathbf{n_i}) = \mathbf{T_i}, p\,[\leq \mid = \mid \geq]\,\hat{\mathbf{P}}_\mathbf{i}$. The bold sysmbols correspond to the actual values of DOF, t-score and reported p-value, respectively. The expression $[\leq \mid = \mid \geq]$ denotes a choice of either one of the three signs: i) less than, ii) equals or iii) greater than, respectively. Finally, $\hat{P}_i$ can be either an exact p-value or a bound on the p-value depending on the $[\leq \mid = \mid \geq]$ sign used [Association, 2009]. We only considered unique p-values from each publication by removing all duplicates of datapoints that have the same values of the triplet $(T_i, n_i, \hat{P}_i)$ as another datapoint in the *same* publication, with the same equality/inequality sign. By inspection, these were always duplicate reports of the same hypothesis test results.

Subsequently, in the reports collected, some of the datapoints were reported with exact p-values (i.e. $t(\mathbf{n_i}) = \mathbf{t_i}, p = \hat{\mathbf{P}}_\mathbf{i}$), and others were reported with either a lower bound or an upper bound (i.e., $t(\mathbf{n_i}) = \mathbf{t_i}, p \geq \hat{\mathbf{P}}_\mathbf{i}$ or $t(\mathbf{n_i}) = \mathbf{t_i}, p \leq \hat{\mathbf{P}}_\mathbf{i}$, respectively). Therefore, in order to perform the analysis as described in the methods section, we recomputed the p-values P_i from the reported t-scores and degrees of freedom using Equation (4.1).

There are a total of 26,119 datapoints in the collected dataset. We used a random 10% : 90% split of the dataset for purposes of exploration (primarily for fixing the null hypothesis $H_0\,(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5)$) and analysis, respectively. The absolute number of datapoints used in the split are 2,614 for exploration and 23,505 for analysis. The analysis portion of the dataset will serve as the dataset D from the methods section, yielding $M = 23{,}505$.

Figure 4.2 depicts the number of occurrences of reported p-values in the exploration part of the dataset in the range $p \in [0.001, 2]$ (note the logarithmic scale on the x axis). An interesting observation to notice from Figure 4.2 is that there is a sudden drop in the number of occurrences of p-values around $p = 0.05$, which is widely used as a significance level for statistical tests. We therefore set $\bar{p}_3 = 0.05$ for the null hypothesis. From the exploration dataset, we set the values $\bar{p}_1 = 0.03, \bar{p}_2 = 0.04, \bar{p}_4 = 0.06$ and $\bar{p}_5 = 0.07$.

Results The null hypothesis H_0 is now set and states that bias against publications with p-values in the range $[0.05, 0.07)$ relative to publications with p-values in the range $[0.03, 0.05)$ doesn't exist. We use the described methods to calculate an upper bound on the probability of observing a dataset least as extreme as the collected dataset D.

For mode detailed insight, we calculate this bound for multiple subsets of the complete dataset, each differing by the choice of limiting the reports by the DOFs to a different range. Moreover, we

[4]The list of crawled journals can be found in Appendix .1.

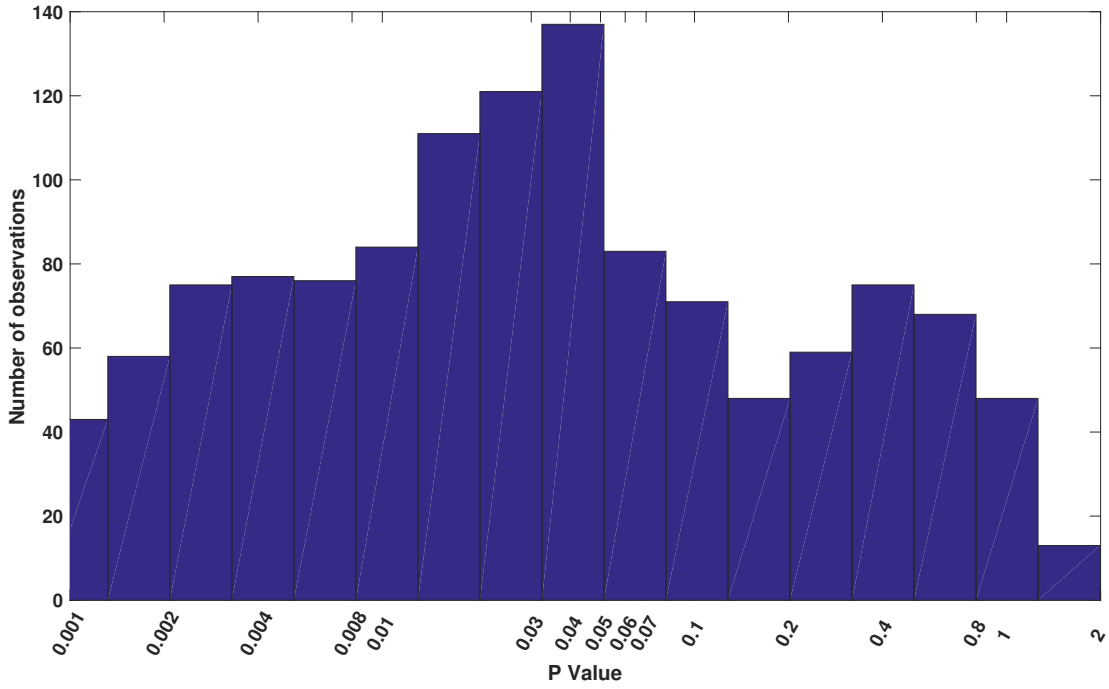


Figure 4.2: Distribution of reported p-values in the exploration dataset.

choose three different mechanisms for partitioning the total set of DOFs, in an effort to understand

1. the loss in tightness in the bound; and
2. the gain in efficiency in computing the bound.

For that purpose, we elected to use the following partitioning mechanisms.

Fine partitioning In this mechanism, given a set, N, of DOFs, we partition it using blocks that are singletons, one per actual observed degree of freedom in N. That is, we create one block per actual degree of freedom. Formally, $\mathcal{N}_p \triangleq \{\{n\} \,|n \in N\}$. As discussed in Section 4.5.2, this yields the tightest bound but the slowest convergence. We will use this bound as a benchmark, measuring the loss in tightness in the other partitioning mechanisms.

Crude partitioning In this mechanisms, given a set, N, of DOFs, we partition it using one block that contains all of N. Formally, $\mathcal{N}_p \triangleq \{N\}$. This partitioning mechanisms, yields the fastest convergence, as discussed in Section 4.5.2. However, this partitioning mechanisms yields the loosest bound, and as such can be used as a measure of the largest loss in tightness, compared to the fine partitioning mechanism.

Reasonable partitioning In this, more gracious, mechanism, given a set, N, of DOFs, we partition it using two types of blocks. The first type of blocks treats the degrees of freedom that are smaller than 30, and is constructed using blocks, each including a range of DOFs of length at most 1. Concretely, each one of these blocks is of the form $([n, n+1) \cap N)$ for $n = \lfloor n_l \rfloor, \ldots, 29$, where $n_l \triangleq \min(N)$. The second type is a single block that contains all DOFs that are at least 30. Concretely, this block is simply $(N \setminus [0, 30))$. Combining these two types of blocks, we formally define $\mathcal{N}_p \triangleq \left(\bigcup_{n=\lfloor n_l \rfloor}^{29} \{[n, n+1) \cap N\} \cup \{N \setminus [0, 30)\}\right) \setminus \{\emptyset\}$.

We subtract the empty set from the union of these two types of blocks for correctness, since we defined partitions to exclude empty sets (think, for example, if there are no datapoints with DOF in the range $[10, 11)$, which would yield $[10, 11) \cap N = \emptyset$, which is not a valid block). This partitioning mechanism is a reasonable middle ground between the fine and crude mechanisms, which has the potential to balance the tightness of the bound and the efficiency of the computation.

The results for different subsets of datapoints, under the different partitioning mechanisms, are compiled in Table 4.2.

Discussing the results As expected, the computations performed with the fine partitioning mechanism were the slowest to converge, and yielded the tightest bounds (smallest). For instance, when we computed the bound for all DOFs of size 2 or larger using fine partitioning (first row in Table 4.2), the computation took more than 24 hours on MATLAB 2016b, running on a laptop with the specs i) 16 GB of memory (1,600 MHz DDR3); ii) 3 GHz Intel core i7 (dual core) central processing unit; iii) 256 kB L2 cache, per core; iv) 4 MB L3 cache; and v) macOS 10.12.5 (Sierra). The computation also utilized MATLAB's parallel computing toolbox, using 2 local workers (one per core).

The same dataset, running under the same conditions described above, but using the crude partitioning mechanism, converged in less than 2 minutes. When using the reasonable partitioning mechanism, the test on same dataset, under the same conditions described above, converged in less than 21 minutes.

It can also be seen that these significant speedups do not come with an expensive price. The tightness of the bound did not suffer a heavy price, as the bound increased from 5×10^{-6} to 6.4×10^{-6} on the same dataset, between the fine and reasonable partitioning mechanisms, respectively. The speedup offered by using the crude partitioning mechanism, compared to the reasonable partitioning mechanism is not as significant, and as such, does not justify the loss in the bound's tightness. Therefore, the choice for reasonable partitioning seems to be a good middle ground, balancing tightness of the bound with the efficiency of computation; therefore, this or similar partitioning mechanisms should be considered when using the toolbox. The conclusions drawn are also applicable for the other subsets described in Table 4.2.

As for the rejection of H_0, in a nutshell, for all subsets of publications that we considered, we reject H_0 since the probability of Type I error is lower than the preset significance level $\alpha = 0.01$ (based on the fine partitioning results). This means that for all subsets of publications that we considered in Table 4.2, we accept the alternative hypothesis that there exists bias against publications with p-values in the range $[0.05, 0.07)$ relative to publications with p-values in the range $[0.03, 0.05)$. This demonstrates the strength of this test. For instance, the same analysis can be performed on a limited set of publications that test the same hypothesis one is trying to study, in order to identify publication bias and move cautiously, if it exists. That said, the probability bound this analysis outputs may be used as a measure of skepticism for drawing conclusions from any set of publications.

Important disclaimer One also needs to be careful when applying this test and interpreting its outcome. The test is designed to identify scenarios of publication bias, not the converse. In other words, this test is *not* designed to prove that a certain set of publication comes without publication bias. In particular, if one calculates the bound this toolbox outputs on a set of publications and it comes, say $p(S \geq s|H_0) \leq 0.8$, it does *not* mean that conclusions can be drawn with

Parameters				Partition Mechanism					
DOF		Statistics		Fine		Reasonable		Crude	
min	max	M	s	θ	p	θ	p	θ	p
2	∞	23,093	634	0.037198	0.000005	0.036294	0.000006	0.019018	0.002130
5	20	8,521	296	0.042914	0.001363	0.043199	0.001365	0.035773	0.004051
5	30	11,965	341	0.036535	0.001495	0.036542	0.001495	0.029690	0.004986
5	35	8,521	296	0.036509	0.000892	0.036071	0.000959	0.029139	0.003629
5	100	8,521	296	0.038230	0.000017	0.038197	0.000018	0.030078	0.000176

Table 4.2: The results using partitions with singleton blocks. Each row describes the results for a subset of the dataset corresponding to the set of DOF stated in the first column.

confidence. This said, this tool may be used as a measure of skepticism, but not as a measure of confidence. Using this analysis to the contrary (i.e., claiming absence of publication bias) is analogous to claiming that a mathematical assertion is incorrect simply because one couldn't prove it by contradiction.

4.7 Summary

Publication bias is a serious problem that distorts the image of science. In its presence, our ability to draw objective conclusions regarding the correctness of scientific hypotheses becomes handicapped. This problem will also affect our ability to correctly identify risk factors of diseases and prevention mechanisms. This, in turn, will unnecessarily delay the realization of a predictive healthcare model. Publication bias, however, is not particular to predictive medicine; it affects all branches of the empirical sciences.

As such, it is important to, first and foremost, eliminate this problem. However, until that happens, we need tools to cope with the problem. One set of tools deals with estimating the level of publication bias in an effort to help correct for it. Unfortunately, these methods suffer from the curse of making unverifiable assumptions. Some of these estimates may be used to detect bias by checking the magnitude of estimated bias and calculating its significance. Again, since the estimates suffer from unverifiable assumptions, any conclusion drawn from them also suffers from the same fate.

In this chapter, we tackled this very problem. We identified the need for a statistical test that can detect publication bias without making unverifiable assumptions. We devised a statistical significance analysis method that, given a dataset of publications (with results from the Student t-test), outputs an upper bound on the probability of observing a dataset at least as extreme as the one in hand. This probability bound can be used to reject the (null) hypothesis that there is no publication bias, and in turn accept the (alternative) hypothesis that the dataset in hand is a result of a biased publication process.

Caution needs to be taken when using this test; this test can be thought of as a gauge for skepticism, not confidence. That is, a high upper bound on the probability of observing a dataset at least as extreme as the one in hand does *not* imply a high level confidence in that the dataset is a product of an unbiased publication process; *nor* does it imply that one may necessarily draw conclusions with high confidence. This test was designed to detect publication bias, not to affirm the lack thereof.

Afterwards, we discussed the implementation details of a MATLAB toolbox that performs this

test on a given dataset. We showed that, if one is willing to pay in the level of tightness of the bound, then one can gain in efficiency (without losing the property that the bound is correct). That is, this does *not* come at the expense of calculating an approximate bound; the bound will remain correct even if the relaxation in tightness is chosen.

Using this implementation, we demonstrated the test on a dataset of 23,505 publications reporting results using the Student t-test from APA journals. In our analysis, we showed that all the subsets considered for this demonstration resulted in the same conclusion of accepting publication bias. That is, for every publications subset we considered, we arrived at the conclusion that these publications are a product of a biased publication process. In the same experiment, we demonstrated the aforementioned trade-off of bound tightness versus computational efficiency. We concluded that some partitioning techniques can be used to drastically speed up computation, without incurring a high loss in the tightness of the bound.

We believe that detecting publication bias must become a standard not only in scientific research, but also in the scientific publishing business. Journals and other publication agencies should conduct regular audits of their substrate of publications and take measures to minimize bias in publications. In addition to their usefulness to researchers, statistical tests that can detect publication bias–like the one presented in this chapter–can help publishers implement the aforementioned audit protocols.

Acknowledgments

Many people have directly or indirectly contributed to this chapter. First and foremost, I would like to thank my dissertation committee members, Professors Ruzena Bajcsy, John Canny and Deirdre Mulligan for their contributions, support, feedback and ideas, which greatly improved the quality of this chapter. In particular, John Canny has been instrumental, every step of the way, in the work presented in this chapter; I learned a lot from him, and for all of that and more, I am in his debt. I am immensely grateful to my good friend Yusuf Erol for all his comments and feedback on many aspects of the work presented in this chapter. I would like to thank Jim Pitman for his feedback and advice on the formal methods presented in this chapter.

This work was supported in part by TRUST, Team for Research in Ubiquitous Secure Technology, which receives funding support for the National Science Foundation (NSF award number CCF-0424422). This manuscript was made possible by Grant Number HHS 90TR0003/01. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the United States Department of Health and Human Services. This work was supported in part by the Center for Long-Term Cybersecurity (CLTC) at UC Berkeley. The views expressed in this paper are those of the authors and do not necessarily represent the official views of the CLTC.

Any errors or mistakes that made it to the final version of this chapter, including typographical ones, are solely my responsibility, not that of any person or entity mentioned above.

Bibliography

Association, A. P. *Publication Manual of the American Psychological Association, 6th Edition.* American Psychological Association (APA), 6th edn., July 2009. ISBN 9781433805615. URL `http://amazon.com/o/ASIN/1433805618/`.

Baker, M. Statisticians issue warning over misuse of P values. *Nature*, vol. 531(7593), 2016. doi: doi:10.1038/nature.2016.19503. URL `http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503`.

Begley, C. G. and Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature*, vol. 483(7391):pp. 531–533, 2012.

Copas, J. What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 162(1):pp. 95–109, 1999.

Coursol, A. and Wagner, E. E. Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, vol. 17(2):pp. 136–137, 1986.

Dickersin, K., Chan, S., Chalmersx, T., Sacks, H., and Smith, H. Publication bias and clinical trials. *Controlled clinical trials*, vol. 8(4):pp. 343–353, 1987.

Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. Publication bias in clinical research. *The Lancet*, vol. 337(8746):pp. 867–872, 1991.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Bmj*, vol. 315(7109):pp. 629–634, 1997.

Hedges, L. V. Modeling publication selection effects in meta-analysis. *Statistical Science*, pp. 246–255, 1992.

Hedges, L. V. and Vevea, J. L. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, vol. 21(4):pp. 299–332, 1996.

Ioannidis, J. P. and Trikalinos, T. A. An exploratory test for an excess of significant findings. *Clinical Trials*, vol. 4(3):pp. 245–253, 2007.

Iyengar, S. and Zhao, P.-L. Maximum likelihood estimation for weighted distributions. *Statistics & Probability Letters*, vol. 21(1):pp. 37–47, 1994.

Johnson, N. and Welch, B. Applications of the non-central t-distribution. *Biometrika*, vol. 31(3/4):pp. 362–389, 1940.

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, vol. 490(7419):pp. 187–191, 2012.

Nelson, N., Rosenthal, R., and Rosnow, R. L. Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, vol. 41(11):p. 1299, 1986.

Peng, R. The reproducibility crisis in science: A statistical counterattack. *Significance*, vol. 12(3):pp. 30–32, 2015.

Pocock, S. J., Hughes, M. D., and Lee, R. J. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *The New England journal of medicine*, vol. 317(7):p. 426, 1987.

Prinz, F., Schlange, T., and Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, vol. 10(9):pp. 712–712, 2011.

Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological bulletin*, vol. 86(3):p. 638, 1979.

Rosenthal, R. and Gaito, J. The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, vol. 55(1):pp. 33–38, 1963.

Rosenthal, R. and Gaito, J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, vol. 15(2):pp. 570–570, 1964.

Smith, M. L. Publication bias and meta-analysis. *Evaluation in Education*, vol. 4:pp. 22–24, 1980.

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908.

Sutton, A. J., Duval, S., Tweedie, R., Abrams, K. R., and Jones, D. R. Empirical assessment of effect of publication bias on meta-analyses. *Bmj*, vol. 320(7249):pp. 1574–1577, 2000.

Terrin, N., Schmid, C. H., Lau, J., and Olkin, I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in medicine*, vol. 22(13):pp. 2113–2126, 2003.

Wadman, M. NIH mulls rules for validating key results. *Nature*, vol. 500(7460):p. 14, 2013.

Wasserstein, R. L. and Lazar, N. A. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, March 2016. doi:10.1080/00031305.2016.1154108. URL `http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf`.

Chapter 5

Final Thoughts

We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time.
– Thomas Stearns Eliot OM, 1943

Currently, the healthcare model in the United States of America (US) has a reactive nature. Simply speaking, we usually seek medical advice or care only *after* we perceive a deterioration in our health status. In other words, we first get sick, then go see a doctor. Instead, it would be desirable to move towards a more proactive healthcare model. A particular model that we examined in this dissertation, is the predictive healthcare model (also known as predictive medicine).

The road to predictive medicine, however, is long. In this dissertation, we have only taken one step forward in this journey. The realization of the predictive healthcare model has the potential to save lives, improve quality of care and reduce costs for many. More work remains to be done. In particular, we need to prioritize the efforts of discovering risk factors of diseases and their prevention. These risk factors can be incorporated in predictive models, which may be used to predict clinical deterioration before it occurs. If the risk of such deterioration is found to be high, medical intervention can be delivered, before the symptoms manifest.

Medical intervention, on the other hand, relies on understanding means of disease prevention. The revelations of disease risk factors and prevention mechanisms come as a result of expensive, longitudinal, empirical research. There is therefore a need to study means to streamline these studies in a manner that reduces their cost. Telemonitoring is one promising technology that not only can reduce the cost of epidemiological studies, but can also serve as an implementation of the predictive healthcare model. Telemonitoring, though, is not the only option; more research is encouraged to further refine existing protocols or create new ones in order to make long, costly, epidemiological studies more feasible.

Whatever the technology may be, ethical thinking has to be involved in the design process of that technology. One burning ethical issue, when it comes to health-related data, is privacy. We presented a semantic model that has the potential to achieve perfect privacy protection, without costing in utility. The research, however, should not stop there. Instead, we should further investigate the special privacy requirements of predictive medicine. Beyond privacy, other fundamental ethical issues need to be addressed. This line of research is long and takes dedication, not only from academics and entrepreneurs, but also from decision and policy makers. For example, health insurers in the US currently don't require their coverage applicants to go through genetic testing or profiling. However, these ideas have floated by insurers. The ethical repercussions of such requirement, for example, are vast.

This is just one example of the many important matters we need to attend to. Unfortunately, policy has consistently lagged behind technological innovation. We argue that this may be dangerous in the realm of predictive medicine. Inherent to this model are predictions that estimate our health status. Whenever predictions become accurate, we start getting the same types of privacy risks we usually consider when disclosing sensitive information. Except in this case, we are not disclosing sensitive information, they are rather *inferred* about us. In simpler terms, we understand, only a little better, the risks involved by disclosing a certain piece of private information. But do we understand the risks involved in disclosing pieces of information that seem innocuous?

For instance, rating movies on movie streaming services, such as Netflix, may seem innocent, in a privacy sense. However, what if these ratings become the input to a very accurate predictor of sexual orientation? These questions have to be asked, and we have to get ahead of them. The traditional approaches to privacy will start to break if we do not tackle the risks from statistical inference.

Going back to health insurers, who in the backbone of their profit analysis, assess the risk of someone needing medical attention in the future. Given that it is acceptable for insurers to profile the risk of different applicants to cost them money, where does the line cross? Are they allowed to use accurate disease predictors to assess that risk as well (since these conditions are not really disclosed, but predicted)? Note that in this scenario, the health insurer tries to assess the risk of, say cancer, from whatever markers they are allowed to use by law. But what if now, their prediction of cancer, based on data that they may use for risk assessment, becomes highly accurate? If they choose to deny insurance to applicants based on very accurate *predictions*, wouldn't that be comparable to denying coverage based on prior medical conditions?

But these issues aren't the only ethical issues involved. We, scientists, have an ethical obligation to represent an uncensored and undistorted image of science and reality. As such, we have an obligation, that has ethical implications, to report about our failed trials and experiments. Although we discussed, in this dissertation, a statistical method that can detect such censorship through the mechanism of bias, this is hardly sufficient. We truly need to eliminate any mechanism of censorship to begin with. We need protocols that would incentivize researchers to disclose all empirical studies they perform (including ones that are viewed as unsuccessful). Devising these protocols, although requires fluency in statistical theory, is a task that we collectively bear on our shoulders; including laymen, the media and policy makers. For instance, media coverage of "sensational" scientific findings has to become responsible reporting so to not distort the public opinion as to what is a true scientific fact versus what is false or unverified. Moreover, publishers need to address this issue as well; but by large, more effort needs to be taken in this area.

In summary, although we have taken a small step forward towards functional predictive medicine, we need to accelerate our collective steps, but without losing our north. We need to never cease to wonder, question, explore and *be skeptical*. Moreover, we need to change the culture of undervaluing the so-called "unsuccessful" experiments for they may teach us more about the truth than what is perceived to be successful (or what is perceived to be the truth). In short, stay thirsty and keep exploring.

Fiat lux!

Glossary

AI Artificial Intelligence. ii

APA American Psychological Association. 104, 105, 113, 117, 138

API application programming interface. 31, 33, 38, 42, 45

BDSG Bundesdatenschutzgesetz. 68

Berkeley Telemonitoring is an effort to study and develop telemonitoring systems. The effort resulted in a framework for mobile health (mHealth)-based telemonitoring designed for Android devices. 11, 13, 20, 22, 23, 27, 30, 33, 36–47, 49–52, 60, 63, 64, 82, 93, 94, *see* telemonitoring & mHealth

BLE Bluetooth Low Energy. 41–43

BMI Body Mass Index. 80, 89, 90, 92, 93

CDC Center for Disease Control and Prevention. 3, 7, 13, 80, 89, 93

CDF cumulative distribution function. 108

CHF congestive heart failure. 4–7, 11, 19–27, 30, 36, 50, 52, 61–64, 66, 74, 93, 94

CLTC Center for Long-Term Cybersecurity. iii, 53, 94, 118

CMS Centers for Medicare & Medicaid Services. 4, 5, 7

COPD chronic obstructive pulmonary disease. 5, 6

DOF Degrees of Freedom. 13, 106, 108–116

ECG electrocardiogram. 31

EE energy expenditures. 21, 23–28, 35, 38, 43, 48, 50, 51

EECS the department of Electrical Engineering and Computer Sciences. ii

EU European Union. 67–71, 73

FTC Federal Trade Commission. 67, 69–71

GDP gross domestic product. 8, 9

GPS Global Positioning System. 27, 43, 44, 50, 73

HART the Human-Assistive Robotic Technologies lab. ii

HHS Department of Health and Human Services. iii, 53, 94, 117

HIPAA Health Insurance Portability and Accountability Act. 32, 36

IoT Internet of Things. 72

mammography is the most common breast cancer screening modality. 2

MATLAB **mat**rix **lab**oratory; a numerical computing programming language developed by MathWorks, Inc. 13, 91, 105, 109, 110, 112, 115, 117

Medicare is a US national social insurance program administered by the federal government since 1965. The program mainly covers individuals that are 65 years old and older; and younger individuals with disabilities. 4–8

mHealth mobile health. 12, 13, 19–22, 27, 30, 33, 35–38, 52

MITM Man-In-The-Middle. 60, 66, 77

NCI National Cancer Institute. 2

NIH National Institutes of Health. 104

NMFF Northwestern Medical Faculty Foundation. 19, 23, 25

NSF National Science Foundation. iii, 53, 94, 117

NYU New York University. 19

oikos is the private sphere of domestic life (greek philosophy; Aristotle). 59

OM The Order of Merit, recognizing distinguished service in the armed forces, science, art, literature, or for the promotion of culture. 121

OOP Object-Oriented Programming. 38

P3P is Privacy Preferences Platform. A specification–endorsed by World Wide Web Consortium (W3C)–for machine-readable privacy policies, designed for Web resources. 67, 71, 73, 75

PDF probability density function. 107

PDI Private Disclosure of Information. 13, 41, 60, 66, 77–82, 88, 90, 91, 93, 94

personalized medicine is a healthcare model under which medical practices and decisions are tailored to the individual patient. 3, 24, *see* precision medicine

PHD ISO/IEEE 11073 Personal Health Device. 42, 43, 50

PMF is Privacy Mapping Function, the sanitization mechanism that is learned and used in Private Disclosure of Information (PDI). 79, 81–88, 90, 91, 93

polis is the public sphere of political activity (greek philosophy; Aristotle). 59

PPACA Patient Protection and Affordable Care Act. 4, 5, 9

precision medicine is a healthcare model under which medical practices and decisions are tailored to the individual patient. 3, 13

predictive healthcare is a healthcare model under which raliable predictions about the risk of clinical deterioration are sought and interventions are performed in cases where this risk is deemed high. 2–5, 7–13, 19, 20, 23, 24, 36, 52, 60, 66, 72, 73, 76, 93, 101, 102, 116, 121, 122, *see* precision medicine & personalized medicine

preventative healthcare is a healthcare model under which actions are sought in order to prevent deteriorations in health before they occur. 3

privacy is the right of an individual or group to keep certain information about themselves from the public sphere. 13, 66

Privacy by Design is a systems design paradigm in which privacy is incorporated in the design phase of the system. 67, 71–73, 75, 76

PSA is prostate-specific antigen. A glycoprotein enzyme often elevated in the presence of prostate cancer. 2

readmission is the act of unplanned rehospitalization after an initial discharge. 4, 5

RunningCoach is a remote monitoring and coaching Android app that was designed on top of the Berkeley Telemonitoring framework. 20, 22, 49–52, 63–65, 93, 94

SDK software development kit. 30, 33

SHARP Strategic Health IT Advanced Research Projects. iii

SQL Structured Query Language. 48

SSL Secure Sockets Layer. 25, 47, *see* TLS

Student t-test is a statistical hypothesis test often used to determine significant difference between two (or more) groups in empirical studies. 13, 101, 103–106, 112, 113, 117

SVM Support Vector Machine. 89, 91

TCP Transmission Control Protocol. 47

telemonitoring is a process in which subjects are remotely monitored for clinical progress. 11–13, 19–30, 33, 35–47, 49, 50, 52, 60–64, 66, 72, 73, 76–81, 86, 87, 102, 121

TLS Transport Layer Security. 47, *see* SSL

TRUST Team for Research in Ubiquitous Secure Technology. iii, 53, 94, 117

Type I error is the probability of incorrectly rejecting a true null hypothesis, the probability of a "false positive". 99–101, 103, 104, 113, 115

US United States of America. 2–5, 7–9, 32, 36, 59, 67–71, 121

USD United States Dollar. 8

W3C World Wide Web Consortium. 67, 71, 75

XML Extensible Markup Language. 67, 71

Bibliography

Chapter 1

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D., Kurillo, G., and Bajcsy, R. Smartphone Based Real-Time Health Monitoring and Intervention. In Khan, S. U., Zomaya, A. Y., and Abbas, A. (eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, chap. ?? Springer, 2017a. In press.

Aranki, D., Kurillo, G., Mani, A., Azar, P., van Gaalen, J., Peng, Q., Nigam, P., Reddy, M. P., Sankavaram, S., Wu, Q., and Bajcsy, R. A telemonitoring framework for android devices. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 282–291. IEEE, 2016a.

Aranki, D., Kurillo, G., Sarver, H., Song, E., Asuncion, C., Serven, L., Balakrishnan, U., and Bajcsy, R. RunningCoach – Cadence-Oriented Training Application for Long-Distance Runners, 2017b. In preperation.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Continuous, real-time, tele-monitoring of patients with chronic heart-failure: lessons learned from a pilot study. In *Proceedings of the 9th International Conference on Body Area Networks*, pp. 135–141. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-Time Tele-Monitoring of Patients with Chronic Heart-Failure Using a Smartphone: Lessons Learned. *IEEE Transactions on Affective Computing*, vol. 7(3):pp. 206–219, July 2016b. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2554118.

Bishop, L., Holmes, B. J., and Kelley, C. M. National consumer health privacy survey 2005. *California HealthCare Foundation, Oakland, CA*, 2005.

Butler, J. and Kalogeropoulos, A. Worsening heart failure hospitalization epidemic. *Journal of the American College of Cardiology*, vol. 52(6):pp. 435–437, 2008.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1968-1978. 1968-1978. CDC WONDER Online Database, compiled from Compressed Mortality File CMF 1968-1988, Series 20, No. 2A, 2000. Accessed: Apr 3, 2017, URL `http://wonder.cdc.gov/cmf-icd8.html`.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1979-1998. 1979-1998. CDC WONDER Online Database, compiled from Compressed Mortality File CMF 1968-1988, Series 20, No. 2A, 2000 and CMF 1989-1998, Series 20, No. 2E, 2003. Accessed: Apr 3, 2017, URL http://wonder.cdc.gov/cmf-icd9.html.

Centers for Disease Control and Prevention. National Center for Health Statistics: Compressed Mortality File 1999-2015. December 1999-2015. CDC WONDER Online Database, compiled from Compressed Mortality File 1999-2015 Series 20 No. 2U, 2016, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed: Apr 3, 2017, URL http://wonder.cdc.gov/cmf-icd10.html.

Centers for Medicare & Medicaid Services. Medicare Ranking for all Short-Stay Hospitals by Discharges Fiscal Year 2005 versus 2004. September 2006. Accessed: Apr 3, 2017, URL https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/SSDischarges0405.pdf.

Centers for Medicare & Medicaid Services. National health expenditures. 2015. Accessed: Mar 2017, URL https://www.cms.gov/Research-Statistics-Data-and-systems/Statistics-Trends-and-reports/NationalHealthExpendData/Downloads/highlights.pdf.

Commonwealth Fund. Why not the best? Results from the national scorecard on US health system performance, 2006. *New York: The Commonwealth Fund*, September 2006. Accessed: Apr 3, 2017, URL http://www.commonwealthfund.org/publications/fund-reports/2006/sep/why-not-the-best--results-from-a-national-scorecard-on-u-s--health-system-performance.

Commonwealth Fund. Why not the best? Results from the national scorecard on US health system performance, 2011. *New York: The Commonwealth Fund*, October 2011. Accessed: Apr 3, 2017, URL http://www.commonwealthfund.org/publications/fund-reports/2011/oct/why-not-the-best-2011.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, vol. 117(6):pp. 743–753, 2008. ISSN 0009-7322. doi:10.1161/CIRCULATIONAHA.107.699579. URL http://circ.ahajournals.org/content/117/6/743.

Egan, R. L. Mammography, an aid to diagnosis of breast carcinoma. *JAMA*, vol. 182(8):pp. 839–843, 1962.

Etzioni, R., Tsodikov, A., Mariotto, A., Szabo, A., Falcon, S., Wegelin, J., Karnofski, K., Gulati, R., Penson, D. F., Feuer, E., et al. Quantifying the role of PSA screening in the US prostate cancer mortality decline. *Cancer Causes & Control*, vol. 19(2):pp. 175–181, 2008.

Gheorghiade, M., Zannad, F., Sopko, G., Klein, L., Piña, I. L., Konstam, M. A., Massie, B. M., Roland, E., Targum, S., Collins, S. P., et al. Acute heart failure syndromes. *Circulation*, vol. 112(25):pp. 3958–3968, 2005.

Giamouzis, G., Kalogeropoulos, A., Georgiopoulou, V. V., Laskar, S., Smith, A. L., Dunbar, S. B., Triposkiadis, F., and Butler, J. Hospitalization epidemic in patients with heart failure: risk factors, risk prediction, knowledge gaps, and future directions. *Journal of Cardiac Failure*, vol. 17(1):pp. 54–75, 2011.

Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., Conover, M., Montague, J., Dillaway, C., Bartczak, K., Ross, J., et al. Hospital-wide (all-condition) 30-day risk-standardized readmission measure. *Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Retrieved September*, vol. 10:p. 2012, 2011.

Hsiao, C.-J. and Hing, E. *Use and characteristics of electronic health record systems among office-based physician practices, United States, 2001-2012*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.

Hussain, M., Al-Haiqi, A., Zaidan, A., Zaidan, B., Kiah, M., Anuar, N. B., and Abdulnabi, M. The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations. *Computer Methods and Programs in Biomedicine*, vol. 122(3):pp. 393–408, 2015.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., and Thun, M. J. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, vol. 58(2):pp. 71–96, 2008.

Jencks, S. F., Williams, M. V., and Coleman, E. A. Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, vol. 360(14):pp. 1418–1428, 2009.

Keenan, P. S., Normand, S.-L. T., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., Schuur, J. D., Stauffer, B. D., Bernheim, S. M., Epstein, A. J., et al. An Administrative Claims Measure Suitable for Profiling Hospital Performance on the Basis of 30-Day All-Cause Readmission Rates Among Patients With Heart Failure. *Circulation: Cardiovascular Quality and Outcomes*, vol. 1(1):pp. 29–37, 2008.

McConnell, M. V., Shcherbina, A., Pavlovic, A., Homburger, J. R., Goldfeder, R. L., Waggot, D., Cho, M. K., Rosenberger, M. E., Haskell, W. L., Myers, J., et al. Feasibility of Obtaining Measures of Lifestyle From a Smartphone App: The MyHeart Counts Cardiovascular Health Study. *Jama cardiology*, vol. 2(1):pp. 67–76, 2017.

National Cancer Insitite. Prostate-Specific Antigen (PSA) Test. 2012. URL `https://www.cancer.gov/types/prostate/psa-fact-sheet`.

National Cancer Insitite. Breast Cancer Screening. 2017. URL `https://www.cancer.gov/types/breast/hp/breast-screening-pdq`.

Obama, B. United states health care reform: Progress to date and next steps. *JAMA*, vol. 316(5):pp. 525–532, 2016. doi:10.1001/jama.2016.9797. URL `http://dx.doi.org/10.1001/jama.2016.9797`.

Shapiro, S., Strax, P., and Venet, L. Evaluation of periodic breast cancer screening with mammography: methodology and early observations. *JAMA*, vol. 195(9):pp. 731–738, 1966.

Shapiro, S., Venet, W., Strax, P., Venet, L., and Roeser, R. Ten-to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*, vol. 69(2):pp. 349–355, 1982.

Thompson, I. M., Goodman, P. J., Tangen, C. M., Lucia, M. S., Miller, G. J., Ford, L. G., Lieber, M. M., Cespedes, R. D., Atkins, J. N., Lippman, S. M., et al. The influence of finasteride on the development of prostate cancer. *New England Journal of Medicine*, vol. 349(3):pp. 215–224, 2003.

US Cancer Statistics Working Group et al. United States Cancer Statistics: 1999–2013 Incidence and Mortality Web-based Report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, 2016. URL `https://nccd.cdc.gov/uscs/`.

US Congress. Patient Protection and Affordable Care Act. *Public Law*, (111-148), 2010.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60(309):pp. 63–69, 1965.

Welch, H. G. and Albertsen, P. C. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *Journal of the National Cancer Institute*, vol. 101(19):pp. 1325–1329, 2009.

Chapter 2

Alshurafa, N., Eastwood, J.-A., Nyamathi, S., Liu, J. J., Xu, W., Ghasemzadeh, H., Pourhomayoun, M., and Sarrafzadeh, M. Improving Compliance in Remote Healthcare Systems Through Smartphone Battery Optimization. *Biomedical and Health Informatics, IEEE Journal of*, vol. 19(1):pp. 57–63, Jan 2015. ISSN 2168-2194. doi:10.1109/JBHI.2014.2329712.

Apple Inc. ResearchKit Programming Guide - Creating Surveys at http://researchkit.org/docs/docs/Survey/CreatingSurveys.html. 2016. Accessed: 09/10/2016.

AppleInsider Staff. Over 10K participants sign up for Stanford medical trial after ResearchKit debut at http://appleinsider.com/articles/15/03/11/over-10k-participants-sign-up-for-stanford-medical-trial-after-researchkit-debut. 2015.

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D., Balakrishnan, U., Sarver, H., Serven, L., Asuncion, C., Du, K., Gruis, C., Peh, G. X., Xiao, Y., and Bajcsy, R. RunningCoach – Cadence Training System for Long-Distance Runners. In *2017 Health-i-Coach – Intelligent Technologies for Coaching in Health*. May 2017a.

Aranki, D., Kurillo, G., and Bajcsy, R. Smartphone Based Real-Time Health Monitoring and Intervention. In Khan, S. U., Zomaya, A. Y., and Abbas, A. (eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, chap. ?? Springer, 2017b. In press.

Aranki, D., Kurillo, G., Mani, A., Azar, P., van Gaalen, J., Peng, Q., Nigam, P., Reddy, M. P., Sankavaram, S., Wu, Q., and Bajcsy, R. A telemonitoring framework for android devices. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 282–291. IEEE, 2016a.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-time Tele-monitoring of Patients with Chronic Heart-Failure Using a Smartphone: Lessons Learned. *IEEE Transactions on Affective Computing*, 2016b. doi:10.1109/taffc.2016.2554118. URL `http://dx.doi.org/10.1109/TAFFC.2016.2554118`.

Asuncion, C., Balakrishnan, U., Sarver, H., Serven, L., and Song, E. *A Telemonitoring Solution to Long-Distance Running Coaching*. Master's thesis, EECS Department, University of California, Berkeley, May 2016. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/`.

Axisa, F., Dittmar, A., and Delhomme, G. Smart clothes for the monitoring in real time and conditions of physiological, emotional and sensorial reactions of human. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 4, pp. 3744–3747. IEEE, 2003.

Azar, P., Mani, A., Peng, Q., and van Gaalen, J. *Expanded Telehealth Platform for Android*. Master's thesis, EECS Department, University of California, Berkeley, May 2015. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2015/`.

Ben-Zeev, D., Schueller, S. M., Begale, M., Duffecy, J., Kane, J. M., and Mohr, D. C. Strategies for mHealth Research: Lessons from 3 Mobile Intervention Studies. *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42:pp. 157–167, 2015. ISSN 0894587X. doi:10.1007/s10488-014-0556-2.

Bloss, R. Wearable sensors bring new benefits to continuous medical monitoring, real time physical activity assessment, baby monitoring and industrial applications. *Sensor Review*, vol. 35(2):pp. 141–145, 2015.

Boulos, M. N. K., Brewer, A. C., Karimkhani, C., Buller, D. B., and Dellavalle, R. P. Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online Journal of Public Health Informatics*, vol. 5(3), 2014.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, vol. 13(3):pp. 1–17, 2011. ISSN 14388871. doi:10.2196/jmir.1838. `arXiv:1011.1669v3`.

Case, M. A., Burwick, H. A., Volpp, K. G., and Patel, M. S. Accuracy of Smartphone Applications and Wearable Devices for Tracking Physical Activity Data. *JAMA*, vol. 313(6):pp. 625–626, 2015.

Chaudhry, S. I., Barton, B., Mattera, J., Spertus, J., and Krumholz, H. M. Randomized trial of telemonitoring to improve heart failure outcomes (Tele-HF): study design. *Journal of cardiac failure*, vol. 13(9):pp. 709–714, 2007.

Chaudhry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., Phillips, C. O., Hodshon, B. V., Cooper, L. S., and Krumholz, H. M. Telemonitoring in patients with heart failure. *New England Journal of Medicine*, vol. 363(24):pp. 2301–2309, 2010.

Chen, C. and Womack, B. Google Reveals Health-Tracking Wristband at `http://www.bloomberg.com/news/articles/2015-06-23/google-developing-health-tracking-wristband-for-health-research`. June 2015. Accessed: 09/10/2016.

Chen, K. Y. and Sun, M. Improving energy expenditure estimation by using a triaxial accelerometer. *Journal of Applied Physiology*, vol. 83(6):pp. 2112–2122, 1997.

Chih, M. Y., Patton, T., McTavish, F. M., Isham, A. J., Judkins-Fisher, C. L., Atwood, A. K., and Gustafson, D. H. Predictive modeling of addiction lapses in a mobile health application. *Journal of Substance Abuse Treatment*, vol. 46(1):pp. 29–35, 2014. ISSN 07405472. doi:10.1016/j.jsat.2013.08.004. NIHMS150003, URL http://dx.doi.org/10.1016/j.jsat.2013.08.004.

Clark, R. A., Inglis, S. C., McAlister, F. A., Cleland, J. G., and Stewart, S. Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. *BMJ*, vol. 334(7600):p. 942, 2007.

Dannecker, K. L., Petro, S. A., Melanson, E. L., and Browning, R. C. Accuracy of fitbit activity monitor to predict energy expenditure with and without classification of activities. *Medicine & Science in Sports & Exercise*, vol. 43(5):p. 62, 2011.

Dhurandhar, N. V., Schoeller, D. A., Brown, A. W., Heymsfield, S. B., Thomas, D. M., Sørensen, T. I., Speakman, J. R., Jeansonne, M. M., and Allison, D. B. Energy balance measurement: when something is not better than nothing. *International Journal of Obesity*, 2014.

Donaire-Gonzalez, D., de Nazelle, A., Seto, E., Mendez, M., Nieuwenhuijsen, M. J., and Jerrett, M. Comparison of physical activity measures using mobile phone-based CalFit and actigraph. *Journal of Medical Internet Research*, vol. 15(6), 2013.

Dwork, C. Differential privacy. In *Automata, Languages and Programming*, pp. 1–12. Springer, 2006.

EMB/11073. ISO/IEEE Health informatics – Personal health device communication Part 00103: Overview. *ISO/IEEE Std 11073-00103:2012*, 2012.

Eng, D. S. and Lee, J. M. The promise and peril of mobile health applications for diabetes and endocrinology. *Pediatric diabetes*, vol. 14(4):pp. 231–238, 2013.

Galbreath, A. D., Krasuski, R. A., Smith, B., Stajduhar, K. C., Kwan, M. D., Ellis, R., and Freeman, G. L. Long-term healthcare and cost outcomes of disease management in a large, randomized, community-based population with heart failure. *Circulation*, vol. 110(23):pp. 3518–3526, 2004.

Giamouzis, G., Mastrogiannis, D., Koutrakis, K., Karayannis, G., Parisis, C., Rountas, C., Adreanides, E., Dafoulas, G. E., Stafylas, P. C., Skoularigis, J., Giacomelli, S., Olivari, Z., and Triposkiadis, F. Telemonitoring in chronic heart failure: a systematic review. *Cardiology Research and Practice*, vol. 2012, 2012.

Hamill, J., Derrick, T. R., and Holt, K. G. Shock attenuation and stride frequency during running. *Human movement science*, vol. 14(1):pp. 45–60, 1995.

Heiderscheit, B. C., Chumanov, E. S., Michalski, M. P., Wille, C. M., and Ryan, M. B. Effects of step rate manipulation on joint mechanics during running. *Medicine and science in sports and exercise*, vol. 43(2):p. 296, 2011.

Hunter, D. L. An Apple a day keeps the research ethics committee away? *Research Ethics*, vol. 11(1):pp. 2–3, 2015. ISSN 1747-0161. doi:10.1177/1747016115585299. URL `http://rea.sagepub.com/lookup/doi/10.1177/1747016115585299`.

Hussain, M., Al-Haiqi, A., Zaidan, A., Zaidan, B., Kiah, M., Anuar, N. B., and Abdulnabi, M. The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations. *Computer Methods and Programs in Biomedicine*, vol. 122(3):pp. 393–408, 2015.

Inglis, S. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. *Journal of Evidence-Based Medicine*, vol. 3(4):pp. 228–228, 2010.

Jardine, J., Fisher, J., and Carrick, B. Apple's ResearchKit: smart data collection for the smartphone era? *Journal of the Royal Society of Medicine*, vol. 108(8):pp. 294–296, 2015. ISSN 0141-0768. doi:10.1177/0141076815600673. URL `http://jrs.sagepub.com/lookup/doi/10.1177/0141076815600673`.

Klaassen, R., op den Akker, R., and op den Akker, H. Feedback Presentation for Mobile Personalised Digital Physical Activity Coaching Platforms. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, pp. 64:1–64:8. ACM, New York, NY, USA, 2013. ISBN 978-1-4503-1973-7. doi:10.1145/2504335.2504404. URL `http://doi.acm.org/10.1145/2504335.2504404`.

Lee, Y.-D. and Chung, W.-Y. Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring. *Sensors and Actuators B: Chemical*, vol. 140(2):pp. 390–395, 2009.

Martin, C. K., Miller, A. C., Thomas, D. M., Champagne, C. M., Han, H., and Church, T. Efficacy of SmartLoss(SM) , a smartphone-based weight loss intervention: Results from a randomized controlled trial. *Obesity*, vol. 23(5):pp. 935–42, 2015. ISSN 1930-739X. doi:10.1002/oby.21063. URL `http://onlinelibrary.wiley.com/doi/10.1002/oby.21063/full`.

Mladenov, M. and Mock, M. A step counter service for Java-enabled devices using a built-in accelerometer. In *Proceedings of the 1st International Workshop on Context-Aware Middleware and Services: Affiliated With the 4th International Conference on Communication System Software and Middleware (COMSWARE 2009)*, pp. 1–5. ACM, 2009.

Ong, M. K., Romano, P. S., Edgington, S., Aronow, H. U., Auerbach, A. D., Black, J. T., De Marco, T., Escarce, J. J., Evangelista, L. S., Hanna, B., et al. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the better effectiveness after transition–heart failure (BEAT-HF) randomized clinical trial. *JAMA internal medicine*, vol. 176(3):pp. 310–318, 2016.

Pande, A., Zeng, Y., Das, A. K., Mohapatra, P., Miyamoto, S., Seto, E., Henricson, E. K., and Han, J. J. Energy expenditure estimation with smartphone body sensors. In *Proc. of the 8th International Conference on Body Area Networks*, pp. 8–14. 2013.

Paré, G., Moqadem, K., Pineau, G., and St-Hilaire, C. Clinical effects of home telemonitoring in the context of diabetes, asthma, heart failure and hypertension: a systematic review. *Journal of Medical Internet Research*, vol. 12(2), 2010.

Park, J.-g., Patel, A., Curtis, D., Teller, S., and Ledlie, J. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 113–122. ACM, 2012.

Patel, M. S., Asch, D. A., and Volpp, K. G. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, vol. 313(5):pp. 459–460, 2015. doi:10.1001/jama.2014.14781. URL `http://dx.doi.org/10.1001/jama.2014.14781`.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *Biomedical Engineering, IEEE Transactions on*, vol. 58(1):pp. 7–11, 2011.

Remme, W. J. and Swedberg, K. Guidelines for the diagnosis and treatment of chronic heart failure. *European heart journal*, vol. 22(17):pp. 1527–1560, 2001.

Rickles, N. M., Svarstad, B. L., Statz-Paynter, J. L., Taylor, L. V., and Kobak, K. A. Pharmacist telemonitoring of antidepressant use: effects on pharmacist–patient collaboration. *Journal of the American Pharmacists Association*, vol. 45(3):pp. 344–353, 2005.

Ritter, S. Apple's Research Kit Development Framework for Iphone Apps Enables Innovative Approaches to Medical Research Data Collection. *Clinical Trials*, vol. 5(2):p. 1000e120, 2015. ISSN 21670870. doi:10.4172/2167-0870.1000e120.

Rowlands, A. V., Eston, R. G., and Tilzey, C. Effect of stride length manipulation on symptoms of exercise-induced muscle damage and the repeated bout effect. *Journal of sports sciences*, vol. 19(5):pp. 333–340, 2001.

Spring, B., Gotsis, M., Paiva, A., and Spruijt-Metz, D. Healthy apps: Mobile devices for continuous monitoring and intervention. *IEEE Pulse*, vol. 4(6):pp. 34–40, 2013. ISSN 21542287. doi:10.1109/MPUL.2013.2279620.

Swedberg, K., Cleland, J., Dargie, H., Drexler, H., Follath, F., Komajda, M., Tavazzi, L., Smiseth, O. A., Gavazzi, A., Haverich, A., et al. Guidelines for the diagnosis and treatment of chronic heart failure: executive summary (update 2005). *European heart journal*, vol. 26(11):pp. 1115–1140, 2005.

Taylor, A. G. The ResearchKit Health Projects. In *Get Fit with Apple Watch*, chap. 8, pp. 111–117. Apress, 2015. ISBN 978-1-4842-1282-0.

Warburton, D. E. R., Nicol, C. W., and Bredin, S. S. D. Health benefits of physical activity: the evidence. *CMAJ : Canadian Medical Association Journal = Journal de l'Association medicale canadienne*, vol. 174(6):pp. 801–9, 2006. ISSN 1488-2329. doi:10.1503/cmaj.051351. `arXiv:1011.1669v3`, URL `http://www.ncbi.nlm.nih.gov/pubmed/16534088`.

Chapter 3

Aranki, D. and Bajcsy, R. Private Disclosure of Information in Health Tele-monitoring. *arXiv preprint arXiv:1504.07313*, 2015.

Aranki, D. and Bajcsy, R. Private Disclosure of Information MATLAB Toolbox at `https://telemonitoring.berkeley.edu/PDI/`. 2016.

Aranki, D., Balakrishnan, U., Sarver, H., Serven, L., Asuncion, C., Du, K., Gruis, C., Peh, G. X., Xiao, Y., and Bajcsy, R. RunningCoach – Cadence Training System for Long-Distance Runners. In *Proceedings of Health-i-Coach '17*. ACM, Barcelona, Spain, May 2017.

Aranki, D., Kurillo, G., Yan, P., Liebovitz, D. M., and Bajcsy, R. Real-time tele-monitoring of patients with chronic heart-failure using a smartphone: Lessons learned. *IEEE Transactions on Affective Computing*, vol. 7(3):pp. 206–219, 2016.

Banzhaf, W., Nordin, P., Keller, R. E., and Francone, F. D. *Genetic programming: An introduction*, vol. 1. Morgan Kaufmann Publishers, Inc., 1998.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2 edn., 2006.

Cranor, L., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J., Schunter, M., Stampley, D. A., and Wenning, R. *The platform for privacy preferences 1.1 (P3P1. 1) specification*. The World Wide Web Consortium (W3C), 1.1 edn., November 2006. URL `https://www.w3.org/TR/P3P11/`.

Department of Justice's Office of Privacy and Civil Liberties (OPCL). Overview of The Privacy Act of 1974. 2015. Accessed: July 2017, URL `https://www.justice.gov/opcl/overview-privacy-act-1974-2015-edition`.

du Pin Calmon, F. and Fawaz, N. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pp. 1401–1408. IEEE, 2012.

Dwork, C. Differential privacy. In *Automata, Languages and Programming*, pp. 1–12. Springer, 2006.

European Parliament. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal L*, vol. 281:pp. 31–50, October 1995. Accessed: July 2017, URL `http://data.europa.eu/eli/dir/1995/46/oj`.

Evfimievski, A., Gehrke, J., and Srikant, R. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 211–222. ACM, 2003.

Jiao, J., Courtade, T., Venkat, K., and Weissman, T. Justification of logarithmic loss via the benefit of side information. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 946–950. IEEE, 2014.

Langheinrich, M. Privacy by design—principles of privacy-aware ubiquitous systems. In *Ubicomp 2001: Ubiquitous Computing*, pp. 273–291. Springer, 2001.

Miller, B., Huang, L., Joseph, A. D., and Tygar, J. D. I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis. In *Privacy Enhancing Technologies: 14th International Symposium, PETS 2014, Amsterdam, The Netherlands, July 16-18, 2014. Proceedings*, pp. 143–163. Springer International Publishing, 2014. ISBN 978-3-319-08506-7. doi: 10.1007/978-3-319-08506-7_8. URL http://dx.doi.org/10.1007/978-3-319-08506-7_8.

Narayanan, A. and Shmatikov, V. How To Break Anonymity of the Netflix Prize Dataset. *CoRR*, vol. abs/cs/0610105, 2006. URL http://arxiv.org/abs/cs/0610105.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE, 2008.

Nissenbaum, H. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

Pitofsky, R., Anthony, S. F., Thompson, M. W., Swindle, O., and Leary, T. B. Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress. 2000. Note: Commissioner Swindle dissented from the report and Commissioner Leary concurred in part and dissented in part.

Pitofsky, R., Azcuenaga, M. L., Anthony, S. F., Thompson, M. W., and Swindle, O. Privacy online: A report to congress. 1998.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, vol. 58(1):pp. 7–11, 2011.

Privacy Act. US Congress. *5 U.S.C.*, (§ 552a), 1974.

Rebollo-Monedero, D., Forne, J., and Domingo-Ferrer, J. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22(11):pp. 1623–1636, 2010.

Reed, I. S. Information theory and privacy in data banks. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, pp. 581–587. ACM, 1973.

Salamatian, S., Zhang, A., du Pin Calmon, F., Bhamidipati, S., Fawaz, N., Kveton, B., Oliveira, P., and Taft, N. How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data. In *GlobalSIP*, pp. 269–272. 2013.

Sankar, L., Rajagopalan, S. R., and Poor, H. V. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, vol. 8(6):pp. 838–852, 2013.

Schaar, P. Privacy by design. *Identity in the Information Society*, vol. 3(2):pp. 267–274, 2010.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10(05):pp. 557–570, 2002.

Tavani, H. T. and Bottis, M. The consent process in medical research involving DNA databanks: some ethical implications and challenges. *ACM SIGCAS Computers and Society*, vol. 40(2):pp. 11–21, 2010.

US Congress. Federal Trade Commission Act. *15 U.S.C.*, (§§ 41-58), 1914.

van Rossum, H. *Privacy enhancing technologies: the path to anonymity*. Registratiekamer, 1995.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60(309):pp. 63–69, 1965.

Warren, S. D. and Brandeis, L. D. The right to privacy. *Harvard Law Review*, pp. 193–220, 1890.

Westin, A. F. *Privacy and Freedom*. Atheneum, New York, 5 edn., 1968.

Westin, A. F. and The Staff of The Center for Social & Legal Research. Bibliography of Surveys of the U.S. Public, 1970-2003. 2003. Accessed through the Wayback Wachine (July 2017), URL `http://www.privacyexchange.org/survey/surveys/surveybibliography603.pdf`.

White, A. M., Matthews, A. R., Snow, K. Z., and Monrose, F. Phonotactic reconstruction of encrypted VoIP conversations: Hookt on fon-iks. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pp. 3–18. IEEE, 2011.

Yamamoto, H. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.). *IEEE Transactions on Information Theory*, vol. 29(6):pp. 918–923, 1983.

Chapter 4

Association, A. P. *Publication Manual of the American Psychological Association, 6th Edition*. American Psychological Association (APA), 6th edn., July 2009. ISBN 9781433805615. URL `http://amazon.com/o/ASIN/1433805618/`.

Baker, M. Statisticians issue warning over misuse of P values. *Nature*, vol. 531(7593), 2016. doi: doi:10.1038/nature.2016.19503. URL `http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503`.

Begley, C. G. and Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature*, vol. 483(7391):pp. 531–533, 2012.

Copas, J. What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 162(1):pp. 95–109, 1999.

Coursol, A. and Wagner, E. E. Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, vol. 17(2):pp. 136–137, 1986.

Dickersin, K., Chan, S., Chalmersx, T., Sacks, H., and Smith, H. Publication bias and clinical trials. *Controlled clinical trials*, vol. 8(4):pp. 343–353, 1987.

Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. Publication bias in clinical research. *The Lancet*, vol. 337(8746):pp. 867–872, 1991.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Bmj*, vol. 315(7109):pp. 629–634, 1997.

Hedges, L. V. Modeling publication selection effects in meta-analysis. *Statistical Science*, pp. 246–255, 1992.

Hedges, L. V. and Vevea, J. L. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, vol. 21(4):pp. 299–332, 1996.

Ioannidis, J. P. and Trikalinos, T. A. An exploratory test for an excess of significant findings. *Clinical Trials*, vol. 4(3):pp. 245–253, 2007.

Iyengar, S. and Zhao, P.-L. Maximum likelihood estimation for weighted distributions. *Statistics & Probability Letters*, vol. 21(1):pp. 37–47, 1994.

Johnson, N. and Welch, B. Applications of the non-central t-distribution. *Biometrika*, vol. 31(3/4):pp. 362–389, 1940.

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, vol. 490(7419):pp. 187–191, 2012.

Nelson, N., Rosenthal, R., and Rosnow, R. L. Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, vol. 41(11):p. 1299, 1986.

Peng, R. The reproducibility crisis in science: A statistical counterattack. *Significance*, vol. 12(3):pp. 30–32, 2015.

Pocock, S. J., Hughes, M. D., and Lee, R. J. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *The New England journal of medicine*, vol. 317(7):p. 426, 1987.

Prinz, F., Schlange, T., and Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, vol. 10(9):pp. 712–712, 2011.

Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological bulletin*, vol. 86(3):p. 638, 1979.

Rosenthal, R. and Gaito, J. The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, vol. 55(1):pp. 33–38, 1963.

Rosenthal, R. and Gaito, J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, vol. 15(2):pp. 570–570, 1964.

Smith, M. L. Publication bias and meta-analysis. *Evaluation in Education*, vol. 4:pp. 22–24, 1980.

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908.

Sutton, A. J., Duval, S., Tweedie, R., Abrams, K. R., and Jones, D. R. Empirical assessment of effect of publication bias on meta-analyses. *Bmj*, vol. 320(7249):pp. 1574–1577, 2000.

Terrin, N., Schmid, C. H., Lau, J., and Olkin, I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in medicine*, vol. 22(13):pp. 2113–2126, 2003.

Wadman, M. NIH mulls rules for validating key results. *Nature*, vol. 500(7460):p. 14, 2013.

Wasserstein, R. L. and Lazar, N. A. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, March 2016. doi:10.1080/00031305.2016.1154108. URL `http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf`.

Appendices

.1 Appendix I – Publication Bias Data

The list of American Psychological Association (APA) journals, which were crawled for the purposes of the experiment in Section 4.6 are listed in the table below.

Journal Code	**Journal Name**
60935	Journal of experimental psychology: Human perception and performance
60934	Journal of Experimental Psychology: Learning, Memory, and Cognition
60945	Journal of Consulting and Clinical Psychology
60909	Behavioral Neuroscience
60948	The Journal of Abnormal Psychology
60937	Journal of Experimental Psychology: General
60946	Journal of Comparative Psychology
60979	Psychological Assessment
60938	Journal of Experimental Psychology: Applied
60943	Journal of Counseling Psychology
60939	Journal of Experimental Psychology: Animal Behavior Processes
60983	Professional Psychology: Research and Practice
60986	Personality Disorders: Theory, Research, and Treatment (2009-)
60929	American Psychologist
60912	Training and Education in Professional Psychology (2006-)
60951	International Journal of Play Therapy
726353	Sport, Exercise, and Performance Psychology (2011-)
60990	Journal Of Psychotherapy Integration
726355	Couple and Family Psychology: Research and Practice (2011-)

Table 1: List of the nineteen journals surveyed from APA. Some journals published their first issue after the year 2002, in which case we mention in parentheses the year of the first issue published by that journal. The journals list is sorted according to the number of publications with experiments surveyed in our analysis.