

**Crowdsourcing Health Discoveries:  
from Anecdotes to Aggregated Self-Experiments**

by

Ian Scott Eslick

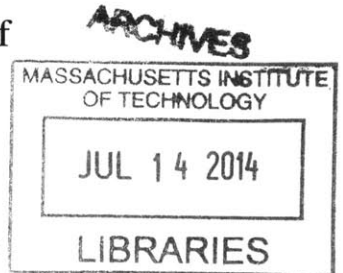
Submitted to the Program in Media Arts and Sciences  
School of Architecture and Planning  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013



© Massachusetts Institute of Technology 2013. All rights reserved.

**Signature redacted**

Author . . . .

Program in Media Arts and Sciences  
School of Architecture and Planning  
August 9th, 2013

**Signature redacted**

Certified by . . . . .

Frank Moss  
Professor of the Practice of Media Arts and Sciences  
Thesis Supervisor

**Signature redacted**

Accepted by . . . . .

.....  
Pattie Maes  
Associate Academic Head  
Program in Media Arts and Sciences



**Crowdsourcing Health Discoveries:  
from Anecdotes to Aggregated Self-Experiments**

by

Ian Scott Eslick

Submitted to the Program in Media Arts and Sciences  
School of Architecture and Planning  
on August 9th, 2013, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

**Abstract**

Nearly one quarter of US adults read patient-generated health information found on blogs, forums and social media; many say they use this information to influence everyday health decisions. Topics of discussion in online forums are often poorly-addressed by existing, clinical research, so a patient's reported experiences are the only evidence. No rigorous methods exist to help patients leverage anecdotal evidence to make better decisions.

This dissertation reports on multiple prototype systems that help patients augment anecdote with data to improve individual decision making, optimize healthcare delivery, and accelerate research. The web-based systems were developed through a multi-year collaboration with individuals, advocacy organizations, healthcare providers, and biomedical researchers. The result of this work is a new scientific model for crowdsourcing health insights: Aggregated Self-Experiments.

The self-experiment, a type of single-subject (n-of-1) trial, formally validates the effectiveness of an intervention on a single person. Aggregated Personal Experiments enables user communities to translate anecdotal correlations into repeatable trials that can validate efficacy in the context of their daily lives. Aggregating the outcomes of multiple trials improves the efficiency of future trials and enables users to prioritize trials for a given condition. Successful outcomes from many patients provide evidence to motivate future clinical research. The model, and the design principles that support it were evaluated through a set of focused user studies, secondary data analyses, and experience with real-world deployments.

Thesis Supervisor: Frank Moss

Title: Professor of the Practice of Media Arts and Sciences



**Crowdsourcing Health Discoveries:  
from Anecdotes to Aggregated Self-Experiments**

by  
**Ian Scott Eslick**

The following people have served on the committee for this thesis:

  
**Signature redacted**

.....  
Frank Moss

Professor of the Practice of Media Arts and Sciences

MIT Media Laboratory

  
**Signature redacted**

.....  
Henry Lieberman

Principal Research Scientist

MIT Media Laboratory

  
**Signature redacted**

.....  
Peter Szolovits

Professor of Computer Science and Engineering

MIT Department of EECS



## Acknowledgments

I would like to thank the members of my committee who have encouraged me in the pursuit of this dissertation. Without the extraordinary support of my advisor, Professor Frank Moss, this work would never have been completed. Professor Peter Szolovits brought a wealth of experience from AI and medicine and has been an invaluable source of encouragement and grounding, particularly for some of the more esoteric ideas I've generated over the years. Dr. Henry Lieberman's creativity, especially in user interfaces and user engagement, has been both genuinely delightful and instructive.

I would also like to thank Linda Peterson and the rest of the MAS administration for their support of my unconventional path to this dissertation.

No large body of work is completed without significant emotional and logistical support. My wife Mary Kelly has been my champion from the day I decided to return to academia. I appreciate all the wisdom she's shared from her own PhD experience. My parents' support of our family during several critical years made sustaining this effort possible through some trying circumstances. My children have been endlessly patient with long nights, weekends at the office, and travel. I look forward to making that up to them.

My mother-in-law Judy Hallagan, herself a nurse and former clinical researcher, has shared with me a unique window into the dysfunction and inhumanity of modern health-care. It is a window none of us wanted to have to look through, but walking the path with her has taught me a great deal while simultaneously fueling my sense of outrage.

Finally, several organizations have been key collaborators and supporters of this work. To Compass Labs, Lybba, NEA, the C3N Project of the Cincinnati Children's Hospital, ImproveCareNow, and the LAM Treatment Alliance: I will be forever grateful for your support.





# Contents

<b>1</b>	<b>Introduction</b>	<b>23</b>
1.1	Crowdsourcing Health Insights . . . . .	25
1.2	Personalized and Evidence-Based Medicine . . . . .	28
1.3	Research Context . . . . .	29
1.4	Experiments in Practice . . . . .	31
1.5	Aggregated Self-Experiments . . . . .	33
1.6	Results . . . . .	35
<b>I</b>	<b>Research Context</b>	<b>37</b>
<b>2</b>	<b>Related Work</b>	<b>39</b>
2.1	Emerging Disciplines . . . . .	39
2.1.1	Peer-to-Peer Healthcare . . . . .	40
2.2	Related Platforms . . . . .	41
2.2.1	Tracking and Aggregation . . . . .	42
2.2.2	Experimentation . . . . .	43
2.2.3	Collective Intelligence . . . . .	44
2.2.4	Quantified Self . . . . .	45
2.2.5	The Learning Healthcare System . . . . .	45
2.3	Knowledge in Patient Forums . . . . .	46

2.4	Design and Analysis of Clinical Trials . . . . .	48
2.5	Single Patient Trials . . . . .	50
2.5.1	Aggregated N-of-1 Trials . . . . .	53
2.5.2	Self-experimentation . . . . .	53
2.5.3	Measurement . . . . .	54
2.6	Bayesian Modeling and Belief Updating . . . . .	55
2.6.1	Belief Update . . . . .	57
<b>3</b>	<b>Health Social Media Forums</b>	<b>59</b>
3.1	Online Health Forums . . . . .	60
3.2	Corpus and Patient Language . . . . .	61
3.2.1	Semantic Knowledge . . . . .	62
3.2.2	Speech Acts . . . . .	64
3.2.3	Anaphoric References and Implicit Context . . . . .	67
3.2.4	Identifying Clinical Terminology . . . . .	68
3.2.5	Topic Models . . . . .	69
3.2.6	Corpus Summary . . . . .	71
3.3	Representation . . . . .	72
3.3.1	Semantic Relations . . . . .	74
3.3.2	Networks of Relations . . . . .	74
3.3.3	Target Representation . . . . .	76
3.4	Information Extraction . . . . .	77
3.4.1	Extraction Example . . . . .	79
3.4.2	Constituent Extraction . . . . .	81
3.4.3	Relation Identification . . . . .	82
3.4.4	Bootstrapping . . . . .	84
3.5	Preliminary Extraction Results . . . . .	85
3.5.1	Constituent Identification . . . . .	85

3.5.2	Relation Extraction . . . . .	87
3.5.3	Discussion . . . . .	87
3.6	Implications . . . . .	90
<b>4</b>	<b>Aggregating Patient Reported Outcomes</b>	<b>91</b>
4.1	Origin of LAMsight . . . . .	92
4.2	Design and Features . . . . .	93
4.3	Community Reception . . . . .	95
4.4	Results . . . . .	97
4.4.1	Patient Questionnaires . . . . .	97
4.4.2	Researcher Questions . . . . .	98
4.4.3	A Study of Pulmonary Function . . . . .	98
4.5	Lessons Learned . . . . .	101
<b>II</b>	<b>Aggregated Self-Experiments</b>	<b>105</b>
<b>5</b>	<b>Extended Example</b>	<b>107</b>
5.1	Case Review . . . . .	107
5.2	Seeking Help Online . . . . .	109
5.3	Personal Experiments . . . . .	110
5.4	Alice’s Trials . . . . .	117
5.4.1	Glycerin and Witch Hazel for Psoriasis . . . . .	117
5.4.2	Turmeric for Psoriasis . . . . .	120
5.4.3	Restriction Diet for Fatigue . . . . .	122
5.4.4	Pagano Diet for Psoriasis . . . . .	123
5.5	Six Months Later . . . . .	126
<b>6</b>	<b>Framework</b>	<b>127</b>

6.1	Hypothesis Space . . . . .	127
6.2	Individual Process Model . . . . .	129
6.2.1	Minimizing Decisions . . . . .	131
6.3	Designing Experiments . . . . .	133
6.3.1	Sharing and Peer Review . . . . .	134
6.3.2	Replication . . . . .	135
6.3.3	Improving Designs . . . . .	136
6.4	Reducing Trial Burden . . . . .	136
6.4.1	Optimizing Sample Size . . . . .	137
6.4.2	Accommodating Confounding Factors . . . . .	138
6.5	Computing the Probability of Success . . . . .	139
6.5.1	Recommending Experiments . . . . .	140
6.6	Deciding to Continue . . . . .	141
6.7	The Population Hypothesis Space . . . . .	141
6.7.1	Predictor Discovery . . . . .	142
6.7.2	Causal Analysis . . . . .	142
6.7.3	Ethical Considerations . . . . .	143
<b>7</b>	<b>Algorithms</b>	<b>145</b>
7.1	The Single-Subject Trial as Decision Aid . . . . .	145
7.2	The Ideal Trial . . . . .	148
7.2.1	Measurement and Variability . . . . .	148
7.2.2	Normalized Treatment Effects . . . . .	150
7.2.3	Planning a Trial . . . . .	151
7.2.4	Updating Beliefs . . . . .	153
7.2.5	Interpreting the Trial Outcome . . . . .	154
7.2.6	Visual Outcomes . . . . .	154
7.3	Working with Real-world Data . . . . .	155

7.3.1	Time series effects . . . . .	156
7.4	Confounding . . . . .	158
7.4.1	Learning from Special Causes . . . . .	159
7.5	Recommending Experiments . . . . .	159
<b>8</b>	<b>Prototype Design</b>	<b>161</b>
8.1	Site Navigation . . . . .	162
8.2	Data Model . . . . .	164
8.2.1	Support for Data Acquisition . . . . .	165
8.2.2	Treatments . . . . .	175
8.2.3	Experiments and Trials . . . . .	176
8.2.4	Miscellaneous . . . . .	180
8.3	Consumer Tracking Workflow . . . . .	180
8.3.1	Tracking . . . . .	181
8.3.2	Interpreting Evidence . . . . .	183
8.4	Consumer Experimentation Workflow . . . . .	184
8.4.1	Trial Actions . . . . .	185
8.4.2	Visualizing Trial Data . . . . .	187
8.4.3	Trial Outcome Reporting . . . . .	188
8.4.4	Successful Treatments and Abandoned Trials . . . . .	188
8.5	Collaboration Features . . . . .	189
8.5.1	Discussion . . . . .	189
8.5.2	Ratings . . . . .	189
8.5.3	Trial Peer Review . . . . .	190
8.5.4	Experiment Peer Review . . . . .	190
8.6	Aggregation Support . . . . .	190
8.7	MyIBD Deployment . . . . .	191
8.8	Technology Architecture . . . . .	195

8.8.1	Server Architecture . . . . .	196
8.8.2	Client Architecture . . . . .	197
8.8.3	Deployment . . . . .	198
8.8.4	Performance . . . . .	201
8.8.5	Lessons Learned . . . . .	201

### **III Evaluation and Discussion 205**

#### **9 Evaluation 207**

9.1	Users Can Design Self-Experiments . . . . .	207
9.1.1	Spontaneous experimentation on social media forums . . . . .	208
9.1.2	Edison: The Experimenter’s Journal . . . . .	210
9.1.3	Experiment Design Survey . . . . .	213
9.1.4	Designs from Personal Experiments . . . . .	215
9.1.5	Discussion . . . . .	216
9.2	Users can Execute and Learn from Self-Experiments . . . . .	216
9.2.1	C3N/ImproveCareNow N-of-1 Experience . . . . .	217
9.2.2	Personal Experiments for Self-Tracking . . . . .	218
9.2.3	Personal Experiments User Study . . . . .	219
9.2.4	Uses of Personal Experiments in the wild . . . . .	227
9.2.5	Discussion . . . . .	228
9.3	Learning from the outcomes of multiple trials . . . . .	229
9.3.1	Analysis of Variance . . . . .	230
9.3.2	Optimizing Sample Size . . . . .	231
9.3.3	Improving Designs . . . . .	233
9.4	Aggregated Self-Experiments will have a Dramatic Impact on Healthcare .	235
9.4.1	Doctor-patient Relationships . . . . .	235

9.4.2	Supporting Hypotheses for Research . . . . .	238
9.4.3	The Learning Health System . . . . .	238
9.5	Observations from Early Prototypes . . . . .	243
<b>10</b>	<b>Future Work</b>	<b>245</b>
10.1	Improving Trial Design . . . . .	245
10.1.1	Exploiting Covariates . . . . .	246
10.1.2	Discovering Carryover Effects . . . . .	246
10.1.3	Modeling Transitions . . . . .	247
10.1.4	Improved Modeling Fidelity . . . . .	247
10.1.5	Retesting the Baseline . . . . .	248
10.1.6	Empirically Chosen Effect Size Categories . . . . .	248
10.2	Alternative Trial Designs . . . . .	248
10.2.1	Short Onset/Offset . . . . .	249
10.2.2	The “Two Armed” Body . . . . .	249
10.2.3	Factorial Design . . . . .	249
10.2.4	Dose-Response Trials . . . . .	250
10.2.5	Dynamic Trials . . . . .	250
10.3	Surrogate Measures . . . . .	251
10.3.1	Surrogate Measures for Clinical Outcomes . . . . .	251
10.3.2	Passive Measures . . . . .	252
10.4	Combination Treatments . . . . .	252
10.5	Accommodating Global Constraints . . . . .	253
10.6	User Interface Elements . . . . .	253
10.6.1	Tagging Taxonomy . . . . .	253
10.6.2	Medical Taxonomy . . . . .	254
10.6.3	Variable Control Limits for Run Charts . . . . .	254
10.7	Integration with Health Social Media Forums . . . . .	254

<b>11 Conclusion</b>	<b>257</b>
<b>A Social Media Content</b>	<b>261</b>
A.1 Edison Experiments . . . . .	261
A.2 Patient Experiments: TalkPsoriasis . . . . .	264
A.2.1 Slippery Elm Treatment . . . . .	264
A.3 Patient Treatment Advice: TalkPsoriasis . . . . .	265
A.4 Blog post by a Personal Experiments User . . . . .	268
<b>B Personal Experiments and MyIBD Catalogs</b>	<b>271</b>
B.1 User Study Experiments . . . . .	271
B.2 Personal Experiments Catalog . . . . .	278
B.2.1 Experiments . . . . .	278
B.2.2 Instruments . . . . .	290



# List of Figures

2-1	Platform Comparison . . . . .	42
3-1	Density of medical content by speech act . . . . .	67
3-2	Manual Spleen Causal Model . . . . .	80
3-3	Conditional Random Field Features . . . . .	82
3-4	A Relation Context Window . . . . .	83
3-5	Precision of Constituent Terms . . . . .	86
3-6	Relation Labels . . . . .	87
4-1	The LAMsight Home Page . . . . .	93
4-2	LAMsight Data Collection . . . . .	94
4-3	LAMsight Data Explorer . . . . .	95
4-4	LAMsight Survey Editor . . . . .	96
4-5	Hormone levels during the female menstrual cycle [Iso09] . . . . .	100
4-6	FEV1/FEV6 with menstrual onset (red) and LH surge (green) . . . . .	102
5-1	Searching for Psoriasis-related Information . . . . .	111
5-2	Measurement, Treatment, and Experiment Views . . . . .	112
5-3	Configuring a Trial . . . . .	114
5-4	User Dashboard . . . . .	115
5-5	Trackers, Trial Control Charts, and Journaling . . . . .	116
5-6	Glycerin and Witch Hazel, Intermediate Trial Results . . . . .	119

6-1	Hypothesis space for a single user . . . . .	128
6-2	Alice’s Scientific Method . . . . .	130
6-3	Mechanisms to Support Individual Decision Making . . . . .	131
6-4	Process Model . . . . .	132
6-5	Community Science . . . . .	133
6-6	Producers and Consumers . . . . .	134
6-7	Common Single-Subject Designs . . . . .	139
7-1	AB Trial with Treatment Effect . . . . .	149
7-2	Cohen’s <i>d</i> Effect Sizes . . . . .	151
7-3	Statistical Model . . . . .	152
8-1	Top Navigation Header . . . . .	162
8-2	Personal Experiments Site Map . . . . .	162
8-3	Search Interface . . . . .	163
8-4	Dashboard Page . . . . .	164
8-5	Chart Review Page . . . . .	165
8-6	High Level Data Schema . . . . .	166
8-7	Measure View . . . . .	167
8-8	Measure Fields . . . . .	167
8-9	Tracker Management . . . . .	169
8-10	Tracker Fields . . . . .	169
8-11	Configuring a Tracker . . . . .	170
8-12	Service Configuration . . . . .	170
8-13	Photo of an SMS “Thread” on an iPhone 5 . . . . .	172
8-14	Treatment View . . . . .	175
8-15	Treatment Fields . . . . .	175
8-16	Experiment View . . . . .	176

8-17	Experiment Fields . . . . .	176
8-18	Advanced Experiment Fields . . . . .	177
8-19	Create a Trial . . . . .	178
8-20	Trial Fields . . . . .	179
8-21	Advanced Trial Fields . . . . .	179
8-22	Search Result Detail . . . . .	181
8-23	My Charts Page . . . . .	182
8-24	Dashboard Trial Widget: Active Trial . . . . .	184
8-25	Timeline Control Chart . . . . .	187
8-26	MyIBD Login Page . . . . .	191
8-27	MyIBD Learning Plan Page . . . . .	192
8-28	MyIBD De-identified Population Browser . . . . .	193
8-29	MyIBD Clinician Chart Review . . . . .	194
8-30	Platform Architecture . . . . .	196
8-31	Deployment Architecture . . . . .	199
9-1	Slippery Elm Trial . . . . .	209
9-2	Edison Experiment Categories . . . . .	212
9-3	Coded Edison Experiments . . . . .	213
9-4	Experiment Design Survey compared to Edison . . . . .	214
9-5	Population Ages . . . . .	220
9-6	Prior Self-Trackers? . . . . .	221
9-7	Population Occupations . . . . .	222
9-8	Selected Interventions . . . . .	222
9-9	Trial of Increase Exercise to Increased Total Sleep . . . . .	224
9-10	Trial of Melatonin Dose impact on Deep Sleep . . . . .	225
9-11	Use of Tea Tree Oil to Reduce Itching from Tinea Versicolor . . . . .	226
9-12	Glycerin and Witch Hazel for Psoriatic Scaling . . . . .	227

9-13 Herbal Liver Treatment for Fatigue in a Psoriasis Patient . . . . .	228
9-14 Deep sleep has a typical Cohen's $d = 0.5$ . . . . .	230
9-15 Deep sleep ratio has a typical Cohen's $d = 1.0$ . . . . .	231
9-16 Study Population Total Sleep . . . . .	232
9-17 Sample size for a two-tailed t-test . . . . .	233
9-18 Simulated Sample Sizes . . . . .	234
9-19 Cystic Fibrosis Symptom Chart . . . . .	236
9-20 MyIBD Ad-Hoc Interventions . . . . .	239
9-21 MyIBD Transplant Recovery Chart . . . . .	241

# List of Tables

3.1	Speech Act Annotation Ontology . . . . .	65
3.2	Medical Topic Modelling . . . . .	70
4.1	Estrogen Study Population . . . . .	100
5.1	Experiment Detail for Glycerin and Witch Hazel . . . . .	118
5.2	Experiment Detail for Turmeric . . . . .	121
5.3	Experiment Detail for the Restriction Diet . . . . .	124
5.4	Experiment Detail for the Pagano Diet . . . . .	125
9.1	User Experiment of L-Tryptophan . . . . .	215



# Chapter 1

## Introduction

“For one-third of U.S. adults, the Internet is a diagnostic tool,” according to a 2013 study of U.S. online health behavior. By the end of 2012, 59% of U.S. adults consumed online health information, 24% of whom sought information from other patients for insight into their own health [FD13a]. Peer advice is largely shared through health social networks [Swa09] as personal anecdotes, such as:

User 1: *I have mild psoriasis on my palms and the bottoms of my feet. I just started using the 50/50 mix of glycerin/witch hazel and felt soothing results right away. ... It took the itch away and really helped with the flakes after just one application!*

User 2: *glycerin witch hazel mixture really didn't help me either. I have better luck with turmeric cream.*

Anecdotes like these rarely convey information about how to use a treatment, how long to wait to see results, the magnitude of effect one can expect to see, or how many people it has helped. Anecdotes do not provide the information necessary to choose among multiple treatment suggestions. More importantly, there is rarely any clinical or comparative effectiveness research available for these kinds of self-care interventions. This dissertation places the tools of science directly into the hands of individuals so they can develop evidence to inform their choices.

Science relies first and foremost on measurement. The rise of cheap mobile devices, networks, computation, and sensors has enabled numerous ways to collect data about our health and everyday lives. To date, such systems have been used primarily to identify correlations, but correlations alone are insufficient for individuals to act confidently.

I present a new framework, Aggregated Self-Experiments, as the missing link between patient anecdotes and self-tracking data. A self-experiment is a documented, reproducible procedure for quantitative measurement of the effect of a treatment while varying exposure to that treatment. The framework estimates the probability that a treatment improved their condition, helping them to make a more informed choice about continuing or abandoning it. The self-experiment is a special case of the n-of-1 clinical trial design [DEG<sup>+</sup>13], but addresses the design challenges that arise when ordinary users design and run experiments.

This dissertation addresses four main questions:

1. Can users design self-experiments?
2. Can users execute and learn from trials of self-experiments?
3. Can we learn from the outcomes of multiple trials?
4. How will this impact the healthcare system?

To evaluate these questions, I developed two prototype web-based tools: one supporting self-experiments for individuals and one for healthcare providers. I report on formative studies, user evaluations, and case studies demonstrating that with appropriate tools, users can design, run, and interpret self-experiments. The dissertation documents the specific features enabling these interactions and provides guidance to designers who wish to implement the framework.

I demonstrate that aggregation of experimental outcomes enables new users to run the same experiment in less time and interpret the results with higher confidence. As more treatments are explored for a given symptom or condition, the framework can identify the more reliable and effective treatments, helping users select among multiple treatment. Choosing better experiments and completing them faster reduces time spent finding a viable



therapy among a large set of suggestions.

The remainder of this chapter provides an overview of the research context, Aggregated Self-Experiments framework, and results. Readers interested in the highlights should find that this chapter satisfies curiosity. The remainder of the dissertation is divided into three parts. Part I introduces relevant academic background and reports on two precursor studies that motivated the concept of aggregating self-experiments. Part II provides the technical details of the Aggregated Self-Experiments framework, starting with an idealized example, providing a formal description of the framework, introducing algorithms to support it, and closing with a detailed review of the prototype systems. Part III answers the four questions posed above by reporting on secondary data analyses, user studies, and experiences from real-world deployments. I conclude by discussing prospects for future research and a final summary of the contributions.

## **1.1 Crowdsourcing Health Insights**

This dissertation focuses on “crowdsourced health insights” or what we can learn directly from the experiences of one another to inform personal decisions and contribute to scientific knowledge. The dominant form of such insight today emerges from peer-to-peer information exchange on health forums or social networking sites. Recent years have seen the emergence of a variety of online services that capture data to augment or complement peer-to-peer sharing. These services can be sub-categorized into social networking, service distribution, and symptom tracking [Swa09]. Social networking refers to patient-to-patient and patient-to-provider interactions around health topics of interest. Consumer services include concierge medicine, spa services, wellness programs, and direct sales of alternative medicines, supplements, and vitamins.

There are several challenges in the prior work that this dissertation addresses. In the online world, people readily consume content, but rarely generate it [Fox09]. When they

do contribute ideas, they rarely ground the debate in verifiable fact (Chapter 3). When people do have facts, they are typically culled from a scientific literature that may not apply to us [GDL<sup>+</sup>09], can be biased [Ioa05], or that they fail to interpret correctly [GGK07]. When facts are not available, people share their experience or stories of other's experiences. Rarely are these stories backed up by evidence. When people do document their experience, they almost always do so using a "correlation as causation" paradigm.

In chapter 3, I identify two major modes of health information sharing: the presentation of "information as fact," often derived from scientific or consumer publications, and the sharing of personal narrative. When debating published fact, we can appeal to the original publication for authority, but we have only personal authority for establishing the validity, utility or generality of a patient narrative.

Engagement in health communities appears to be correlated with how well those services satisfy the information-seeking goals of the user, in preference to other inducements such as emotional or social support [Nam11]. Unfortunately, getting what you want is not the same as getting what you need. Patients who make decisions based on recommendations and data collected online confront a multitude of information-seeking and interpretation biases that can interfere with effective decision making [LC07] [KBK08] [Sto00]. These biases, and a lack of systematic feedback on the effectiveness of narrative information, may in part explain the lack of strong outcome data to date.

Combating bias in discovery and inference is nothing new; the rise of empiricism in the 17th century [Bac90] [Des99] was driven in large part by a desire of the thinkers of the day to escape from "intellectual idolatry" and put inquiry into the natural world on a solid footing. Over the past four hundred years, we have honed and institutionalized a rigorous hypothetical-deductive methodology for falsifying concrete hypotheses. However, the apparent success of this methodology, and its specific embodiment in the form of statistical null-hypothesis testing, is often resistant to alternative methods of building evidence in support of decision making [Kru13] [LPD<sup>+</sup>11]

Cheap sensing, communications, and computing make it feasible to accumulate far more information about individual, day-to-day experience than was possible even a decade ago [Swa12a]. The potential for coupling patient narrative and longitudinal sensor data is exemplified by the Quantified Self (QS) movement [Wol10]. QS consists of thousands of people that assemble in local “meetups” to share stories about what they measured, how they measured it, and what they learned. This movement is one highlight in a larger movement of empowered individuals attempting to utilize tools once restricted to institutional science to gain insight into their own condition and environment.

Symptom tracking is a rapidly expanding scope of activity ranging from fertility and athletic performance to chronic disease management. Increasingly, cheap sensor-based devices are used to capture real-time information from users to be presented via web-based or mobile device visualizations to help inform and guide future behavior. This idea of using online collaboration and captured data to inform our behavior is the heart of the challenge: how effective are the actions taken by individuals in improving outcomes? While there is good evidence that minimal harm is being caused by online health information [Eys04] [Fox09], there is limited evidence thus far that online health information and social interaction improves outcomes [Eys04] [PW10]. What is lacking for effective crowdsourced health insight is a methodology to enable individuals to effectively document their experiences for one another and establish how to satisfy the goals of people seeking health insights from their peers.

When asked about this apparent methodological gap, the founder of The Quantified Self remarked that methodology is a missing element in the movement towards making use of personal data. He said that he has heard only limited discussion of methodology for experimentation over hundreds of presentations [Wol11]. The “what I learned” component of QS presentations almost exclusively reports observed correlations, not confirmation of causality. Seth Roberts, one of the few trained experimentalists in the community to apply controls to personal data [Rob10], remarked that he has tried to teach experimental rigor

to hobbyists and that the complex tradeoffs involved in designing formal experiments appeared to be beyond both the reach and patience of most people, even dedicated hobbyists [Rob12].

## 1.2 Personalized and Evidence-Based Medicine

To develop a methodology for crowdsourcing health insight, it is prudent to examine models developed by institutional medical science. Unfortunately, the state of medical evidence and its translation into practice leaves much to be desired [Ins01]. The challenge of modern medicine is highly structural, consisting of institutional, political, regulatory, financial, and technical barriers [BNW08].

There is a profound human and economic demand to bring better methodology to everyday health decisions. In the US we spend over \$2T per year on standard healthcare. Over 30% of the population uses complementary and alternative therapy [SL11] and we spend over \$25 billion dollars on dietary supplements [Rep09]. Traditional healthcare has significant challenges with translating traditional research into better care and a majority of consumer investment in alternative treatments remains largely uninformed by empirical research, representing a potentially enormous opportunity cost for both individuals and society. This cost could be dramatically reduced if individuals had a low-cost, low-effort means to establish whether a given intervention was effective. The primary goal of the work reported here is to help an individual make better personal healthcare decisions.

Guidance to physicians comes primarily from the outcomes of randomized controlled trials published in peer-reviewed journals. These trials are expensive and time-consuming. The organizations sponsoring these trials, by necessity, are conservative in their aims; trials that cannot provide a return on investment are unlikely to be run, leaving many important questions unasked and negative results unpublished [Ioa05]. Moreover, even when clinical trial data exists, the translation of that trial into clinical practice is sorely lacking

[Gol12]. Heterogeneity of response to a drug is often higher in practice than in carefully controlled clinical trial settings [Rot05] [GDL<sup>+</sup>09]. One executive estimated that over 90% of common medications are efficacious in less than 50% of the treated population [Con03]. Predictors of for whom a treatment works are difficult to identify and rarely reported in the results of RCTs. It then takes an average of 17 years for RCT results to filter into clinical practice . Once a treatment has become part of standard practice, 50% of the time, the best treatments are not prescribed and 50% of correct prescriptions are not taken correctly [Ins01].

The “Learning Health Care System” applies quality improvement techniques at the institutional level, accelerates sharing of research and practice information, and encourages patient-participation in governance[Ins01]. This emerging model of healthcare exhibits many of the properties of interest: more autonomy for individuals, focuses on patient experience and quality of life, and real-time learning at the point-of-care.

### **1.3 Research Context**

The work in this thesis was performed within a model of action research [Hin08]. I worked closely with several advocacy organizations (LAM Treatment Alliance and the American Cancer Online Resources), online user communities (TalkPsoriasis and the LAM Foundation Listserv), and small user populations drawn from the QS community. I was also a member of a project team inside a prototype learning health system centered at the Cincinnati Children’s Hospital and Medical Center (CCHMC). The Collaborative Chronic Care Network Project (C3N Project) [MPS13] is the research arm of a Quality Improvement (QI) network called ImproveCareNow [WM11]. ImproveCareNow is a non-profit organization coordinating collaboration processes and information sharing across 35+ pediatric gastroenterology centers involved in the treatment of Inflammatory Bowel Disease (IBD). The C3N Project evaluates system-level interventions across this multi-center network.

In concert with the LAM Treatment Alliance I talked with researchers, clinicians, patients, and other advocates regarding the process underlying the discovery of medical treatments as well as patient discovery and exchange of treatments and self-care. We identified several opportunities for leveraging insights from patients to accelerate biomedical discovery, translation, utilization, and self-care. I engaged in a data-mining analysis of LAM and cancer-related forums to start to identify what we could glean from these resources (see Chapter 3).

Between 2007 and 2013 I developed and managed a platform called LAMsight for the LAM community [TDSM06] that enabled patients to generate background and longitudinal surveys that documented their experiences or generated data for researchers. A novel exploration feature was provided that enabled patients to visualize population data. We used this platform to run a population study to gather daily data from patients to illuminate an undocumented self-report phenomenon reported in Chapter 4.

A derivative of this platform, The International LAM Registry [NEC<sup>+</sup>10], was developed to help LAM researchers aggregate data across several dozen global registries. This project developed some novel approaches to collaboration around publication and co-authorship rights, and strategies for creating engagement through enabling researchers to answer questions about sub-populations for whom no single registry had sufficient numbers of patients.

During these projects, I encountered the Quantified Self movement and the C3N Project. The quantified self is a hobbyist movement that explores the use of measurement of daily life to gain insights. The C3N Project and the ImproveCareNow network are two of the more progressive organizations involving patients in the improvement of care delivery. These two projects are as close as any today to a true Learning Healthcare System; they provided a perfect incubator to evaluate new methods of patient-contributed discoveries at a faster pace than was possible in the prior context of biomedical research.

How do the experiences of LAMsight translate to improving care delivery? Is there a

model of patient-driven discovery that translates naturally back into the healthcare context? The experience of these three projects clearly implied that more rigorous documentation is required to make effective use of a patient's lived experience, but finding the right incentives was challenging. LAMsight and similar sites indicate that existing efforts to capture longitudinal observational data about outcomes will be difficult to scale.

The question becomes how to provide sufficient incentive for users to engage for a sufficient period of time, gain maximal insight from a limited window of insight into each user's life, and combat traditional sources of bias. National surveys of patients indicate that most people trust the medical system, and the Quantified Self movement demonstrated that it was becoming increasingly easy to gather information about what happened to patients between provider visits. The C3N Project was intensely interested in the utilization of data about patients to improve care. What if providers prescribe data collection to patients instead of just drugs? Would that improve the quality, adherence, and sustainability of longitudinal patient data?

This still doesn't address the issues of data quality. The issues of confidence led to the consideration of experimental methods at the level of individual patients. Experiments are finite tests of a hypothesis about symptoms or treatment-outcome cause and effect that provide the user direct value and feedback. Experiments by definition document treatment, outcome, and confounding factors to tell us something about a given user's response at a given point in time. Because the user has a direct incentive to see an experiment to its conclusion, it is likely that we will gain more value from each experimental interaction.

## **1.4 Experiments in Practice**

One system-level intervention being explored by the C3N Project was a process for performing single-subject, or "n-of-1," trials. A n-of-1 trial is a formal experimental model for characterizing the change in a single patient in response to one or more interventions

at different points in time. These trials help clinicians evaluate a patient’s response to a therapy when that therapy has a high heterogeneity of response or is a therapy about which little is known, such as many complementary and alternative therapies.

The conventional form of a n-of-1 trial in medicine is a randomized, double-blind, often placebo-controlled, multiple-crossover trial design that uses the patient’s past measurements as an experimental control condition [GSA<sup>+</sup>88] [DEG<sup>+</sup>13]. This design works well when the treatment effect is transitory and the patient returns to their untreated baseline state after the treatment effect has “washed out” of the patient. This pattern characterizes treatments for many chronic disease, lifestyle, and exercise-related questions. A wide variety of designs is available to assess outcomes, from the multiple-crossover design to an interrupted time-series design which simply evaluates the pre-and-post treatment condition of the patient to evaluate the impact of the treatment. The use of n-of-1 trials in medical practice is experiencing a resurgence [LPD<sup>+</sup>11] [Ber10] [CHS<sup>+</sup>12], but there is only limited work examining the use of this methodology in the hands of autonomous users.

Researchers have also explored aggregating multiple n-of-1 trials to assess population means and comparative treatment effects [ZSM<sup>+</sup>97] [ZRS10] [NMS<sup>+</sup>10]. This early work illustrates ways in which meaningful inference can be made from the data collected from multiple, independent, methodologically identical n-of-1 trials. The work reported here identifies a set of data aggregation operations that can improve the individual’s use of the n-of-1 trial to assess the efficacy of treatments in their own life.

The loss of information on how individuals actually respond to treatments represents an enormous hole in our healthcare system. While many clinicians may have personal experience with variable response to a disease or the effects of complementary and alternative therapies, the quality of advice is directly proportional to that clinician’s personal experience. The outcomes of these interventions are neither characterized nor systematically shared, leaving the individual patient experience to the vagaries of their clinician’s specific experiences. It is this gap that drives many patients out of the healthcare system and to



the internet, something that should become unnecessary in a mature learning health system that “drive(s) the process of discovery as a natural outgrowth of patient care.” [SSSM]

## 1.5 Aggregated Self-Experiments

Combining the above observations, I developed a framework, Aggregated Self-Experiments, that provides a stronger scientific basis for individual or provider-patient collaborative decision-making in the context of online information sharing and self-tracking. The framework consists of a methodology, a mathematical formalism, a set of inference algorithms, a knowledge representation, and a set of design principles to guide the development of tools that support health related decisions. It was conceived to bridge a gap between the hierarchy of established medical knowledge, most principally the results of randomized, controlled clinical trials, and the ad-hoc process employed by individuals, and often by care professionals, to make decisions about health. Moreover, the model integrates seamlessly with emerging models of provider care and the learning healthcare system.

To evaluate the framework, I designed and implemented two web-based platforms, PersonalExperiments.org and MyIBD, for self-tracking, data sharing, and crowdsourced authoring and execution of self-experiments. These platforms were made available to an open online population, a recruited set of patients with a skin disorder, and a set of users interested in sleep and productivity optimization, as well as deployed within a pediatric medical practice. I present the central design decisions and the implication of specific embodiments on aggregation of experimental outcomes. I further show through mathematical modeling how past experimental outcomes can optimize the selection and execution of individual experiments.

Within the framework, a user seeking a treatment for a health condition decides which of many possible experiments to run, executes a trial, decides whether to incorporate the treatment, and iterates until a successful trial occurs. The framework emphasizes mini-

mization of the total time spent by users to discover one or more treatments they can and will sustain. Aggregated Self-Experiments succeeds when a user runs an experiment that demonstrates improvement in a chosen outcome, and when that user continues the treatment for an extended period beyond the treatment. This is a weaker aim than traditional meta-analysis of trial outcomes, and these trials more closely resemble pragmatic trials than clinical trials.

As such the constraints on this framework differ significantly from those applied to traditional clinical trials as well as traditional n-of-1 trials. Trying out treatments takes place in the noisy environment of everyday patient life, and many of the measures are themselves subjective in nature. Ideally collecting data would be an ongoing and lifelong practice, but sustaining such measurements long term is a high burden for most people. Structuring captured user experience as a sequence of trials and an optimization over that sequence has several benefits. First, trials structure the user experience, providing a focused, short-term point of engagement.

Secondly, each trial's dataset can power several forms of aggregate analysis. For example, each trial provides insight into the behavior of the individual user, as well as the population of users. If trying a second treatment for the same outcome variable, we can use knowledge of the variability of the measurement in the first trial to reduce the necessary sample size of the second. Across many users, we can estimate the average variance and start with a reduced, but not optimal sample size.

Trial outcomes across patients represent replications of an experiment that can also be aggregated. For example, knowledge of the prevalence of successful outcomes can help users select which treatment to try first. Personalization of this prevalence further optimizes the probability of a successful outcome. All of these optimization activities try to reduce burden without compromising the validity of an individual experiment. The scientific validity of the framework rests solely on the quality of the individual experimental designs. The framework provides support for peer review of experimental designs and

individual trials.

This iterative model of recommendation plus self-experimental verification will help address problems where professional medical advice is not available or easily applied. Ideally, the framework described here will converge with related efforts to make n-of-1 trials an extension of standard clinical practice [MPS13] [LPD<sup>+</sup>11] [CHS<sup>+</sup>12] and should generalize to other domains such as classroom-based or online education, management practices, personal productivity, and public policy.

## 1.6 Results

PersonalExperiments.org is the first platform to implement a personalized model of experimentation that enables individuals to act as scientists. User response to the concept of experimentation was strong with users who learned from their individual experiments whether or not it was successful. Further, 95% of interviewed study participants expressed a desire to continue experimenting on the site after the user study.

The framework supports explicit documentation of an experiment. This enables replication of the experiment. Repeated trials of an experiment provides many benefits. I show that:

- Users can replicate an experiment with their own personal trial,
- Replications within a person increases their confidence that a treatment causes a change in outcome,
- Replications across people train models that make it possible to run faster trials,
- Population outcomes can be used to prioritize among multiple experiments for the same outcome or condition, and
- Engaging with an experiment educates users about the process of experimentation.

More importantly, the framework generates data that is useful for healthcare professionals and consistent with trends to develop a Learning Healthcare System that learns

from interactions that happen in the course of ordinary care, and now, daily life. My work with healthcare providers have demonstrated significant conceptual impact and practical adoption. Specifically, I've shown that:

- Patients can use data to present their case and argument to practitioners,
- Patients and caregivers can design and engage in ad-hoc trials of therapy using their everyday experience,
- Biomedical researchers can be motivated by patient-reported population data to consider new hypotheses, and
- Researchers, and increasingly provider organizations, want to explore rigorous models for learning from data collection from patients between office visits.

For developers of patient-facing self-tracking tools, I've identified a set of design principles which should govern the development of tools that engage users through lightweight experimentation:

- Decomposition of experiments into a space of treatment-variable-design,
- Strategies to prefer characterization and long-term adaptation to the conventional approach of rigorous up-front control,
- Algorithms for computing the parameters of practical trials, and
- Explicit feedback loops that improve experiments and parameter estimation over time.

I've also identified a broad set of topics for future research in methodology, statistical modeling, the design of real-world measures, and user interfaces.

This dissertation asserts that Aggregated Self-Experiments, and other approaches built on the concept of lightweight, end-user experimentation, will help self-tracking and social health companies make their data more actionable. Experimental outcomes collected at scale will be a key enabler for the future of personal and biomedical discovery and learning from patient experience at the “point of care” will accelerate the transition to a Learning Health System.

# **Part I**

## **Research Context**



# Chapter 2

## Related Work

Aggregated Self-Experiments was developed in the context of four emerging areas of research related to health and healthcare: peer-to-peer healthcare, self-quantification, collective intelligence, and the learning healthcare system. This chapter introduces these movements along with a brief summary of some of the academic disciplines relevant to the algorithms presented in later chapters. These include the design and analysis of clinical trials, particularly where  $n=1$ , and Bayesian inference techniques.

### 2.1 Emerging Disciplines

As previewed in the prior chapter, users increasingly connect with one another online for reasons beyond emotional support. They discuss advice from clinicians, share their personal experiences, and provide and receive advice from one another. A growing portion of this advice comes in the form of suggestions for self-treatment and self-care. This advice is based primarily on the advice-giver's personal experience or a limited form of word-of-mouth and not what works for the average person or any specific person asking a question. While there is only limited evidence of harm [FD13a], a growing number of companies and researchers are working to understand how patients can play a more direct role in the generation, utilization, and dissemination of health-related knowledge.

In parallel to the verbal and written exchanges, a new phenomenon coined alternatively as self-tracking or self-quantification is experiencing a dramatic growth [Swa09] [FD13b]. Self-quantification refers to the systematic recording of structured data about oneself [Wol10]. This phenomenon ranges from paper journaling and excel spreadsheets to ecosystems of devices and electronic services that track and report detailed health metrics such as sub-second variations in heart rate, skin conductance, heat flux, and mechanical power generated.

The emerging discipline of collective intelligence identifies the conditions under which collections of non-experts, such as individual users of healthcare, are able to exhibit competence approaching or exceeding that of experts. Prediction markets, search engines, and recommendation engines are some of the more well known examples, but many forms of collaborative processes and data aggregation methods are being investigated.

The use of data and past evidence in the day-to-day practice of medicine is undergoing a significant upheaval. Over the past 10 years, some leaders in medicine and healthcare have been calling for the development of a "Learning Healthcare System", characterized by the inclusion of patients as leaders in the healthcare system, rigorous use of transparent measurement and quality science to improve outcomes, and the accumulation of evidence from the point-of-care to improve individual care provider decisions.

Aggregated Self-Experiments is a framework inspired by and suitable for use in a learning healthcare system context, but focused towards serving the needs of users who are not currently empowered by the medical system and turn instead to peer-to-peer healthcare. The framework proposed in this dissertation brings the tools and techniques of self-quantification and collective intelligence to address problems unique to this setting.

### **2.1.1 Peer-to-Peer Healthcare**

A collection of emerging, web-based platforms seek to aggregate patient health data such as symptoms and medication use [NEC<sup>+</sup>10] [BLH10]. Other sites elicit direct opinions



from patient populations to accumulate ratings about treatment efficacy . A few sites such as Genomera are starting to propose novel methods for running population trials online [WVM11] [Swa12c]. All of these sites focus on serving the patient through a combination of visualization and presentation of aggregate statistics; they claim that reproduction of selected clinical trials validate the quality and utility of their aggregate data.

Two major problems plague existing approaches to using patient self-report. First, sites collecting longitudinal data from patient populations do not characterize self-report quality or variation. Self-report bias and inconsistency can be a significant confounding factor in the association of symptoms and conditions with treatments. Second, none of the current sites provide actionable knowledge; they simply present the population data and leave decision-making entirely to the individual.

Patient anecdotes are the dominant source of treatment hypotheses and actionable advice supporting patient decision making outside the healthcare environment. Patients often post questions that are not the subject of clinical research, such as managing treatment side effects, or exploring treatments that are entirely unknown to practitioners and for which no a-priori empirical evidence exists.

## **2.2 Related Platforms**

Aggregated Self-Experiments fills a hole in a landscape of existing platforms which range in data quality from anecdotal platforms, such as various social media platforms, to quasi-experimental such as Edison [Cor]. Existing platforms also vary in the degree to which they produce insights that are personally relevant to a user who contributes content. Figure 2-1 illustrates some of the more well-known platforms developed to help crowdsource or personalize health insights. See Swan for comprehensive review of the early landscape [Swa09].

The earliest communities to aggregate health insights were online forums such as In-

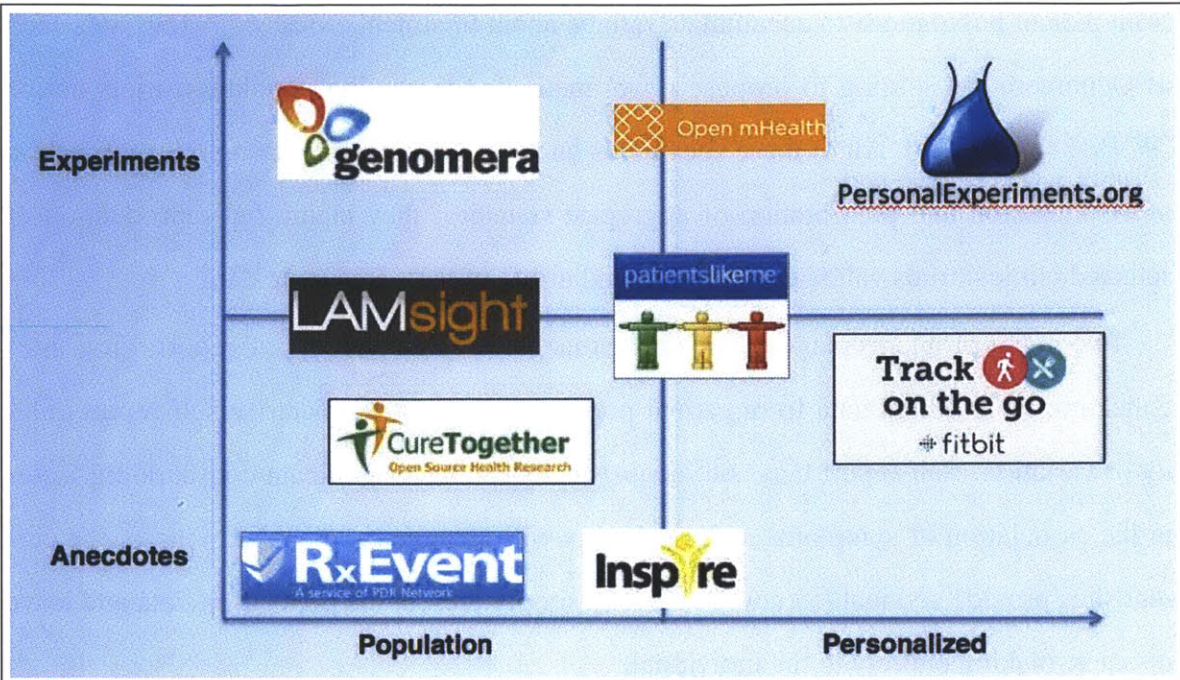


Figure 2-1: Platform Comparison

spire.com and Websphere [Swa09] where patients share freely with one another, often behind privacy barriers. Other sites provide adverse event reporting, typically for physicians, but also some for patients. Patient reported data is anecdotal in nature, although the interaction enabled by forums make them moderately personalized.

### 2.2.1 Tracking and Aggregation

A second generation of services allow users to track symptoms longitudinally, or provide facilities for reporting opinions in a more structured fashion so they can be easily aggregated across a population

- PatientsLikeMe [PW10] is probably the longest running and most well known of data-driven sites for patients and have pioneered many advances in this area. They have built a medical taxonomy for a wide range of conditions and allow users to report their symptoms and compare themselves to other patients “like them”. Thus there are both personal rewards and population benefits to data collection on their

site.

- LAMsight is an early site I developed to allow disease groups to aggregate their data around questions and symptoms of interest to them (in opposition to the taxonomy driven approach of PatientsLikeMe). The site and experience with it is reported in Chapter 4.
- CureTogether.com is a site that allows users to record a diagnosed condition and treatments taken for that condition along with a confidence in how well the treatment helped that condition (a 1-5 scale). The results, however, are population level only; it provides no personalized advice.<sup>1</sup>
- FitBit is both an accelerometer-driven device and a data tracking and visualization service tied to it. They allow you to record daily summaries of passively tracked activity and sleep along with elicited information such as dietary intake. They are highly personalized, and they rely on data, but the resulting insights are limited to correlations. It is hard to know the cause and effect underlying fluctuations in the data.

## 2.2.2 Experimentation

A few sites and initiatives are emerging to explore self-experimentation, predominantly from the perspective of the Quantified Self [Wol10].

- “Edison: The Experimenter’s Journal”: Edison offered semi-structured experiment templates, but was shut down after 3 years due to a lack of uptake. It allowed users to document their personal trials, but did not support their execution of it other than as a place to record the results, journal style.

---

<sup>1</sup>I received a dataset from the owners to evaluate the prospects for generating recommendations for new treatments from patterns of old treatments. Problems of sparsity in the dataset I was given made this challenging, but a simple Slope-1 recommender did a good job of predicting user ratings of held-out treatments. A clustering analysis would be an interesting follow up exercise.

- StudyCure.com: this is a closed beta site that appears inactive by the time of this writing, but had a very simple structured model of experimentation: If I do X, then I'll achieve Y, as measured by Z. Their user interface is a much simpler model of experimentation, but one that hides perhaps too much of the true complexity of real world experiments.
- Open mHealth: this initiative was started to create new standards for exchanging tracking data and running large collections of n-of-1 experiments. It produces standards and reference designs, but their reference trial platform was not available for review at the time of this writing [ES10] [CHS<sup>+</sup>12].

### **2.2.3 Collective Intelligence**

The aforementioned trends have virtually removed barriers for individuals to participate in analysis, publication, and collaborative communication around areas of scientific interest. This, coupled with a substantial global middle class and educated lower class with the liberty to engage one another in topics of interest, are giving rise to new models of creativity and discovery on an almost annual basis. As of the writing of this thesis, the notion of crowd-funding, a follow-on to micro-lending, is transforming the world of finance. The rise of contests to elicit expert contributions on challenge problems and prediction market are entering mainstream use. For the first time in human history, nearly anyone on the planet can theoretically play a critical role in any aspect of scientific discovery or practical problem solving.

Collective Intelligence has been well treated by several prior works including “Wisdom of the Crowds” by Surowiecki and “Democratizing Innovation” by von Hippel [Sur05] [vH05].

The conditions required to enable collective intelligence are as follows.

- Independence: each contributor provides their input without reference to others,

avoiding bias.

- Diversity: each contributor has private information about an aspect of the overall phenomenon.
- Decentralization: user are able to draw on local knowledge.
- Aggregation: a mechanism exists to combine the diverse, independent, decentralized contributions.

von Hippel notes that that lead users can be innovators and play a central role in evolving a new product to fit the unanticipated needs they identify [vH09]. The emergence of hashtag use on Twitter is a simple example of this phenomenon, organizing by topic using the hashtag was not a designed-in function of the Twitter platform [Par11]. Aggregated Self-Experiments draws conceptually from both of the referenced works.

#### **2.2.4 Quantified Self**

The explosion in wireless networking, smart mobile devices, and cheap sensors is lowering the costs of accessing many of the tools of medical science that have historically been restricted to wealthy experimenters and professional institutions. A moment of hobbyists called the Quantified Self has been exploring concepts of self-improvement through data and taking early steps towards a new kind of "citizen science". The same tools are being explored by the healthcare establishment in a movement called mHealth. Both groups are exploring the question of how these new tools change the processes we use improve our own lives, or to deliver better care to others.

#### **2.2.5 The Learning Healthcare System**

The rise of the "e-patient" is characterized by individual users advocating for access to their health care information and to play a more central role in the healthcare process. This

movement has been paralleled by the establishment's call for a radical transformation of medical care into what the Institute of Medicine calls a "Learning Health System" [Ins01], which embraces and transcends the better-known trends such as personalized medicine, evidence-based medicine, quality improvement, and comparative effectiveness research – all of which heavily emphasize empirical research and data-driven methods to improve patient outcomes and quality of life while reducing the cost for the entire system. One of the central tenants of the learning health system is patient involvement and empowerment, something often talked about and rarely embraced by traditional healthcare institutions. Systems are emerging at the forefront of this movement that facilitate groups of patients to act together to support one another and to inform medical care.

## **2.3 Knowledge in Patient Forums**

Patient forums contain a wealth of information about patient's lived experiences and broad opinions about the efficacy of treatments. The state of the art in helping users find information stored in narrative form on forums is tagging, ranking, and keyword search. These techniques help users find content, but what other methods might be employed to document the knowledge embedded in these patient narratives?

Extracting semantic knowledge from text has generated an immense body of work from POS tagging and named entity recognition to parsing and conversion to logical form. This section introduces specific prior work that emphasizes the analysis of patient language for language use (terms and short phrases), dialog speech acts, summarization, semantic category recognition, and relation identification and extraction.

Recent work in the medical literature discusses alignment and overlap of causal patient language with terminology from formal resources such as the UMLS Metathesaurus [Ano09]. This analysis has been done for patient-entered taxonomic information [SW08], patient queries to information systems [Zie03], and patient support forums [SRR05]. These

results all identify a significant body of terminology that does not map cleanly onto formal medical concepts, and yet contain unique language relevant to patients experience that is informative about a disease [WGA07] [KSDK08].

Speech act analysis for e-mail is a generalization of linguistic speech acts [Jur02] and discourse structure [GS86]. The aim is to automatically classify the function or intention of a given message, for example as part of a structured online dialog. Prior work has investigated student project discussions [KCF<sup>+</sup>06] or task and workflow management [CCM04]. Speech acts can be used to characterize the social dynamics of a forum and possibly serve as priors on the type of content that may be embedded within any given high-level speech act.

The identification of key sentences within a message for purpose of summarization was developed by [WM04] to capture the essence of an ongoing decision-making discourse. Cong [CWL<sup>+</sup>08] discusses the difficulty of identifying structured question and answer content, citing challenges of both syntax and language use. They introduce a graph-based algorithm for extracting, ranking, and linking question and answer sentences.

The above techniques all operate on documents; treating sentences, paragraphs or messages as bags and sequences of words and using labeled training sets to train classifiers that cluster or classify these documents. Semantic information extraction work has been performed both for general natural language, as well as in messaging domains such as e-mail. Open domain named entity extraction, event extraction, and other phrase level semantic tagging techniques have been applied to social communications such as e-mail. Rich, sentence level techniques analyze parse structure, head verbs, and nouns to identify commands and domain-specific information. These techniques are typically dependent on manually-annotated datasets, an expensive and time-consuming process. There has been work in bootstrapping, a methodology for boosting extraction performance using a labeled dataset and a large amount of unlabeled data (discussed further in section 3.4).

Specific work on semantic information extraction for the medical domain focused largely

on analyzing formal medical records such as discharge summaries which are typically more constrained than arbitrary patient language. In [SHSU06], small subsets of the UMLS [Ano09] semantic hierarchy are commonly used as a target representation. The work of chapter 3 used a link-grammar parser and the existing UMLS knowledge base to generate features for a support-vector machine. It focuses on a small subset of semantic categories from the UMLS ontology: Disease, Treatments, Substances, Dosages, Symptoms, Tests, Results, and Practitioners. The output of the system is a category label for input n-grams that represent words or clauses from the summary.

## **2.4 Design and Analysis of Clinical Trials**

Medicine is a discipline that lies at the intersection of biology, psychology, and sociology which suffers tremendously from the inherent complexity and diversity of human life. Models developed in medicine lean heavily on the smoothing effects of large numbers and descriptions of average effects. The struggle to move from population-level medicine to personalized medicine is at its heart a struggle to develop models that balance representational complexity with the inherent heterogeneity and dynamism of populations.

Medical care is concerned with the arts of diagnosis and treatment. Medical science provides tools, techniques, and experimental evidence in support of these two activities. In fact, two primary intellectual activities drive all discovery and care in medical science: classification and assessing the effects of interventions on groups of individuals within a class. We can model a human organism as a hidden feature vector with a multitude of possible labels. The art and science of diagnosis takes manifest symptoms of an individual and attempts to infer the maximum likelihood of the feature vector. Given a concrete diagnosis, an action is chosen to effect problematic symptoms. Clinical trials investigate the average response of a population to an intervention conditioned on a disease. In our desire to make sense of an uncertain world, we tend to interpret the results of diagnosis and



outcomes in highly discrete terms.

In the "hard sciences" such as physics, chemistry, and materials, models are relatively easy to test under laboratory conditions and yield insights with universal implications. Relatively simple models such as Gauss's law in physics can be used to describe a wide variety of physical process and have astounding explanatory and predictive powers. Unfortunately, there are many scientific disciplines that investigate phenomena for which no simple, broadly applicable models have, or are likely to be found. Biology and psychology suffer from this limitation and practitioners in these disciplines often suffer from "physics envy", a propensity to seek models that are too simple to be of much use explaining or predicting the behavior of the system at hand [Min07]. Stated simply, most of the biological universe will not yield to elegant description.

Clinical research is a discipline for establishing the causal consequence of applying a treating to population of individuals. Clinical trials involve a structured approach to the intervention such that the resulting data provide evidence of the causal impact of the treatment on outcomes. choose an endpoint, or an empirical measurement that characterizes the medical status of the population. The goal of the trial is to evaluate whether a treatment, and only that treatment, causes a change in the endpoint. Endpoints are measurable variables such as mortality, risk ratio, symptom change, or a change in a laboratory-produced value.

The "gold standard" of clinical research is the double-blind, placebo-controlled, randomized clinical trial. The emphasis in these trial is on absolute control over any influence that might confound the detection of a causal association between treatment and outcome. The major features of these trials are as follows.

- **Blinding:** double blinded trials are trials where neither the people running the trial, nor the trial subjects, are aware of who is getting a treatment and who is not. This was added to clinical trials because the expectations of not just the subject, but also the investigator, can lead to systematic bias in outcome measurements. Psychology

is a powerful force in clinical trials.

- Placebo controls: another testament to the power of psychology is that the expectation of a treatment effect can create that effect in the absence of any causative agent. In fact, I can give a group a sugar pill for their headache telling them it's a sugar pill, and still most will improve.
- Randomization: in any population, there are random effects on the outcome caused by variation within the population (gender, age, etc). There are also inherent measurement variation or unexpected influences that come up in everyday life. With a large enough population, randomly assigning members to a control group (with placebo) or a treatment group, randomizes the unpredictable influences across the two arms. Sizing a trial to ensure balance of the factors is one of the great challenges of clinical research.

Blinding and placebo control are relatively easy to accomplish with drugs, where sugar pills are an easy placebo. It is much harder to do either with, for example, diet.

## **2.5 Single Patient Trials**

Single patient trials are a form of experimental trial that takes place withing a single patient. These are commonly called an N-of-1 trials which means a sample size of 1 [DEG<sup>+</sup>13]. These trials arose in the field of psychology as a way to be more formal about characterizing the response of an individual to highly customized therapy. In the 1980's Guyatt applied the same techniques to single patient trials in biomedicine as a way to address the heterogeneity of treatment effect. In the decades since then several centers have reported on n-of-1 trial centers with mostly positive results [Lar93]. However, most trial centers are shut down after the grants run out. The value of running trials at scale is not sufficiently established that the costs can be justified by mainstream practitioners.

The goal of an N-of-1 trial is not to establish generalizable, causal relationship elucidated by traditional RCT designs, but the selection of a treatment for an individual patient with the pragmatic goal of improving their care. Researchers, most notably Deborah Zucker and Christopher Schmid have explored the use of hierarchical Bayesian techniques to estimate population effects from single-subject trial designs to overcome challenges in certain classes of RCT designs [ZSM<sup>+</sup>97] [ZRS10] [KDK<sup>+</sup>].

This setting has distinct advantages over RCTs that simplify their analysis.

- The primary goal requires only internal validity,
- Intra-patient variation is often much lower than inter-patient variation, and
- Meaningful effect sizes are typically larger than the target of population trials.

In the last two decades single-subject designs have gained acceptance as a valid method for identifying medical interventions that yield the best health outcomes for specific individuals [Ber10]. These techniques are particularly valuable when within-subject variation is much less than among-subject variation. A number of research organizations are actively looking to integrate the single-subject model into the regular practice of clinical medicine [MPS13]. However, single-subject designs are only applicable when certain conditions hold [GSA<sup>+</sup>88] such as:

- **A baseline state.** For the experimental model to work, the patient must have a consistent and measurable baseline state amenable to treatment through behavior, diet or other changes. The hundreds of eligible conditions range from depression and athletic performance to overweight/obesity and chronic disease management.
- **Short treatment effect onset and offset.** Practical experiments require a relatively short time period between onset of treatment and observation of effect. Further, to enable repeated trials, the effect of the treatment needs to decline relatively quickly. I focus on conditions where onset and offset range from hours to one week.

- **Available instrumentation.** To measure treatment effect, we need reliable instruments for measuring outcome and confounder variables. The professional and hobbyist worlds have created a wide variety of both subjective and objective instruments, many of which do not require clinical visits.
- **Measurable confounding factors.** Experimental outcomes may be difficult to establish when confounding factors that influence the outcome measure are present. Confounding can be managed through large sample sizes and randomization (what large clinical trials do), but also through adjustment (measuring and adjusting for confounding factors). Experimental designs must make trade-offs between design complexity and the number of repeated trials necessary to accommodate confounding factors.
- **Avoids adverse effects.** The community can identify treatments that have short or long-term risk for the individual and encourage consultation or monitoring by a professional. Adverse event reporting ensures that the process is responsive to previously unknown risks.

The original N-of-1 trial design introduced by Guyatt is a randomized, two-armed, multiple-crossover study within a single patient [GSA<sup>+</sup>88]. A series of treatment periods is undertaken, each period consisting of a treated and untreated period. The choice of period is randomized via coin flip. They may also be blinded or double-blinded. Blinding is accomplished by having treatment assignment performed by an independent pharmacist when treating with drugs.

Treatment-baseline pairs are continued until sufficient evidence is attained for a meaningful result, either via visual inspection or via a statistical test. Guyatt proposes the use of Student's one-sided t-test, noting that the choice of expected effect size introduces some bias into the experiment, but that it may be OK if effects below the target size are not clinically relevant or of interest to the patient.

Ostensibly the goal of a trial is to generalize to the population of which the trial represents a test. Interestingly, the Guyatt approach did not strictly follow a rigorous Null-Hypothesis Significant Testing (NHST) that dominated the social sciences at the time. Strict methods are often not required because we are only interested in effect that are larger than the noise caused by confounding or other factors. In the field of psychology where n-of-1 trial originated, the use of NHST as the gold standard test of a theory is being hotly debated [Kru13].

### **2.5.1 Aggregated N-of-1 Trials**

Meta-analysis in traditional clinical science aggregates multiple population trials to characterize unexpected variation among similar trials or to improve estimates of the true population effect size. Similarly, in the single-subject literature, meta-analysis techniques are used to improve within-subject effect estimation over repeated trials or to estimate population-level effects [ZSM<sup>+</sup>97] [VdN07] [NMS<sup>+</sup>10]. These same hierarchical modeling methods can be applied to the outcomes of self-experiments.

### **2.5.2 Self-experimentation**

One solution to the problem of interpreting raw, longitudinal self-report data is structuring the collection of that data into a sequence of outcomes of one or more self-experiments. A self-experiment quantifies anecdotes and formalizes the conditions under which data was collected, removing many sources of potential noise and bias. A self-experiment is also proscriptive; running it increases the confidence of the individual as to whether a treatment positively influences their symptoms or not. A patient can demonstrate, to themselves and others, that their anecdotes describe likely causal effects, and not happenstance.

Self-experiments typically lack all the controls of single-subject trials. For example, treatment blinding and placebo control are difficult to apply to self-experiments. However, if the goal is pragmatic improvement of a condition, the self-experiment can be highly

informative. Unlike traditional trials, patients can repeat the intervention over long time periods to verify the repeatability of an effect.<sup>2</sup>

### 2.5.3 Measurement

Medical science applies statistical analysis to measurements taken from our bodies, minds, and environments. By observing and recording these measures, we can be disciplined in our hypotheses, and by applying experimental controls, we can isolate aggregate causal relationships between interventions and measurements. However, the design of measures and even the choice of measures have significant implications for the utility of the knowledge extracted. Two concerns are presented in this chapter:

1. Are we measuring the right phenomenon?
2. Are we measuring it in the right way?

For example, trials of statins measure blood cholesterol as a proxy for heart attack risk. Researchers choose to measure and experiment on cholesterol rather than directly on heart attacks for both practical and theoretical reasons. Practically, trials that look at heart attacks as their endpoint variable would take far too long. Theoretically, the epidemiological evidence suggests that blood cholesterol is associated with heart attack risk, and some drugs that lower cholesterol have reduced heart attack risk. Therefore, the community believes that new compounds can use blood cholesterol as a “proxy variable” for heart attack risk. Reducing one reduces the other, all things being equal. Proxy measurements are useful in that they may have faster response, thus reducing the cost or time it takes to evaluate an intervention. Unfortunately, they are not always perfect as multi-causality can lead to unanticipated violations of the proxy expectation [JRGLL08].

---

<sup>2</sup>Interestingly, the self-experiments in my research won't suffer from the under-reporting bias that regularly occurs in peer-reviewed literature [DAA<sup>+</sup>08] [Ioa08]. For some problems a series of self-experiments may outperform clinical research in both efficiency and quality

Under the rubric of multi-causality, it is unlikely that cholesterol is the only factor playing into heart attack risk. In fact, it may be neither necessary nor sufficient a condition in the physiology of heart attacks. Drugs that reduce cholesterol may have no impact on heart disease. This example illustrates the potential dangers in choosing the wrong measurement [LLS11].

Ignoring the problems with blood cholesterol as a variable, the measurement of free cholesterol itself is reasonably accurate if taken in a standard way (such as 12 hour fast, same time of day, subject has a consistent diet, etc).

Any system that supports self-experimentation has to take into account the mechanisms of *control* that reduce the variability of a measurement, or a *model* of the variation that occurs under ordinary conditions.

## 2.6 Bayesian Modeling and Belief Updating

Bayesian analysis of data have been around in one form or another for nearly 250 years. The posthumous paper by Reverend Bayes in 1763 provides the first published account of reasoning from data to parameters. For the next 200 years, the refinements and extensions of this idea were known as "inverse probability" [Fie06]. The modern formalism for Bayes' Theorem is:

$$P(\theta_i|D) = \frac{P(D|\theta_i)P(\theta_i)}{\sum_j P(D|\theta_j)P(\theta_j)} \quad (2.1)$$

Informally, this relation states that the probability of a parameter given data is equal to the normalized product of the probability of the data given that parameter and the marginal probability of the parameter. The probability relations in this equality have canonical roles summarized as: **posterior** = **likelihood** × **prior**. The original idea of Bayes' was to update beliefs about a parameter given an observation of data.

The denominator in equation 2.1 is a normalizing expression, and for purposes of con-

ceptual clarity can be ignored as the concrete probabilities after a set of updates can be identified by ensuring that the integral over all parameter values sums to 1. This is often notated by the prefix term  $1/Z$  used once below but excluded from future expressions.

The use of these techniques was limited due to the computational burden of belief update and, after the 1920's, the success of "frequentist" techniques such as null-hypothesis significance testing . The development of Markov-Chain Monte Carlo methods in the 1980's dramatically lowered the computational cost of computing posterior distributions, especially when the likelihood function is non-integrable or the structure of the parameter space is complicated, such as with hierarchical Bayesian models [KF09].

The crucial differences between the Bayesian and more commonly used frequentist approaches is firstly, one of interpretation. Probabilities under the frequentist interpretation refer to the proportion of *experimental outcomes* under the assumption that an experiment is repeated, identically, many times. Bayesian probabilities represent a belief about the actual parameters of a system. The second crucial difference is that any change to the experimental design, such as a review of the data prior to completion, changes the sample space of all possible outcomes and it requires sophisticated methods to track, for example, the impact on Type I error [Kru13]. To the extent that the relevant features of the experimental condition are well-represented by the Bayesian model, the Bayesian interpretation is much more flexible.

Flexibility here refers to the ability to directly compute probabilities of interest, and the invariance of these probabilities to observations of in-process data. Trials based on Bayesian interpretations can be adjusted on the fly based on data observed. For example, if baseline sample variance is higher than predicted, additional data can be collected to increase the power of the measurement for the expected effect. An effect that is larger than expected (ignoring confounding) can allow the experimenter to terminate the experiment early. Because the fundamental inference is belief updating and not null-hypothesis testing, any amount of acquired data can be useful for secondary analysis, such as computing the



population variance across many users of the same outcome measure.

The Bayesian approach to analyzing evidence in an experimental context is based on a commitment to incorporate assumptions into the mathematical model of the experiment. The frequentist approach, in contrast, relies on extrinsic factors enforced by an experimental design. One critique of the Bayesian interpretation is when priors are estimated using a subjective procedure by experts or the statistician. There is an objectivist school of Bayesian thought that utilizes "uninformed" prior distributions which assign probability such that no specific commitment is made aside from the structure of the distribution of the data. A uniform assignment of probability over all feasible values is one such example. Uninformed priors are no more subjective than assumptions about the statistical properties of sampled data in frequentist experiments. Of course, the debate between these two schools of thought has been fierce and will not be revisited here. The choice of Bayesian mathematics here is primarily governed not by philosophy, but by the desire to perform a richer set of inferences in response to acquired data than is easily afforded by the frequentist framework.

### 2.6.1 Belief Update

The use of a prior belief immediately suggests step-wise updates of a posterior distribution by recurrence. The posterior of step one is the prior of step 2, representing the original uninformed prior updated by the evidence of step 1. At any step, the probability distribution over parameter values is a product of the original prior and evidence collected to date. For example:

$$\begin{aligned}P(\theta_1|D_1) &= \frac{1}{Z}P(D_1|\theta_0)P(\theta_0) \\P(\theta_2|D_2, D_1) &= \frac{1}{Z}P(D_2|\theta_1)P(\theta_1) \\P(\theta_2|D_2, D_1) &= \frac{1}{Z}P(D_2|\theta_1)P(D_1|\theta_0)P(\theta_0)\end{aligned}$$

In the limit of a large amount of data, the influence of the original prior becomes insignificant. However, the prior is critically important to the interpretation of a single experiment. The role of prior probabilities on interpretation of evidence is itself a sub-science in statistics called model selection, averaging, or checking. These terms refer to formal methods for validating the choice of prior or accommodating inherent uncertainty in the choice of a prior distribution. Adaptation is done by changing the prior or averaging over multiple priors.

# Chapter 3

## Health Social Media Forums

This chapter presents the results of my early investigation into the nature of patient exchange in online forums. The ultimate aim of the work was to investigate how to extract actionable knowledge about treatment outcomes from existing user-provided forum content. The research also characterizes patient language use, the distribution of speech acts within forums, and the forms of propositional knowledge that can be extracted from such resources.

Mining techniques such as these can be used to semi-automatically populate a list of treatments used by users, user opinions about conventional treatment side effects, and the range of symptoms of concern to most patients. The resulting taxonomy can seed resources such as Personal Experiments. Aggregate statistics would ideally provide subjective priors for experiments of treatment-symptom pairs. This is presented as relevant but optional background for the reader, but these techniques were used to inform the primary work of this dissertation and not implemented directly for the Personal Experiments or MyIBD prototype.

### 3.1 Online Health Forums

Online forums, mailing lists and other forms of social media are a rich, but varied source of information about their topic of interest. Health forums and mailing lists in particular are critical tools for patient support, enabling the sharing of experiences, information and provision of support [Swa09]. The knowledge conveyed in such forums can cover everything from reviews of specific service providers to specific ideas about about causes or treatment.

Prior work has established the value and/or promise of patient interaction and self-report for improving the professional's understanding of the disease [LPFH04], for improving standards of care, and adverse events [WSW08]. Information about patient language use and it's relationship to formal medical terminology can improve patients' interaction with one another and with the medical literature [SRR05].

A patient's model of a disease is very different than a clinician's. Capturing an intuitive model of disease etiology and relating it back to an established scientific model of the disease may yield a significant influence on doctor/patient communication and improvements in patient decision making.

Further, patients may share information about alternative treatments or self-care techniques that could be of benefit to one another. Unfortunately, forums are organized around specific discussions, sometimes extending over time but rarely related to other, earlier discussions. There is little work to date that effectively aggregates discussion on a topic across all topics within a single forum, let alone the many fragmented forums that often exist around disease groups, particularly for rare diseases.

Prior research into extracting knowledge from social media has shown great promise at the level of individual question answering or the large scale statistics of search engines [LLS11], however only limited attention has been paid to the aggregation of simple semantic knowledge from these kinds of real-world resources. This chapter characterizes the knowledge captured by several health forums, identifies techniques for extracting and aggregating this data, and motivates the use of longitudinal data collection to ground narrative

discourse that is the primary contribution of this dissertation.

## 3.2 Corpus and Patient Language

A detailed analysis of the LAM Foundation Listserv [LAM09] was performed. The listserv consists of individuals with the rare disease Lymphangiomyomatosis (LAM). LAM is a complex, multi-system disorder that strikes women in their childbearing years. It is characterized by the slow proliferation of smooth muscle tissue in the lungs, kidneys and brain. LAM is debilitating, and ultimately fatal, through destruction of lung tissue and vascular activity which reducing oxygen intake into the body. Decline and mortality rates are highly variable; one observational study illustrated a 10-year survival rate of 85% [TDSM06], however some patients may lose lung function in just a few years from diagnosis. Regardless of the absolute rate of decline, LAM has significant impacts on lifestyle by reducing activity levels, often requiring bulky supplemental oxygen.

A simple web scraper was built to extract message content from the public interface of the list serve archive. Pre-processing included stripping off headers, attachments, embedded html and most inline forwarded text to avoid duplication. There remains some duplication of content from message replies, but less than a few percent of the total. The threading structure of the e-mail was not well represented by the listserv archival system so reconstruction of reply-to chains was not attempted in this analysis.

An initial characterization of the LAM corpus was performed by a random sampling of several hundred sentences manually categorized into specific speech acts. For the 30% of messages containing information directly relevant to disease processes such as symptoms, treatments, or medical events, a manual phrase-level annotation was done to identify both syntactic contexts and lexical terms associated with disease and patient lifestyle relevant knowledge.

### 3.2.1 Semantic Knowledge

Before analyzing the linguistic and discourse structure of the corpus, a few examples from the text that illustrate the nature of the knowledge embedded in it will be informative. The examples also help to identify what features of the text will make it either feasible, or difficult, to automatically extract. Much of the knowledge identified here will be too detailed to easily represent, or too difficult to extract with the techniques described later. The purpose of this section is to identify the opportunities.

Quite often the desired semantics will be directly expressed by a key verb phrase such as “helps” or “can do” as below:

*”A part of going to Pulm. Rehab helps increase lung fuction, maybe your lung fuction is still good enough that you don’t need it yet, also when you are ready for transplant you need to keep your muscle strong so that you have a good recovery. Right now you may have to exercise on your own, do you have a tread mill, or just walking will help.”*

Here the sentence *Pulm Rehab helps increase lung function* is a clear surface expression of a concept which is both definitional and actionable. While the concept *when you are ready for transplant you need to keep your muscle strong* is complex, the assertion is that if this is true, it will cause a *good recovery*. Implicitly we see that this recovery is referring to recovery from a specific event, namely transplant surgery. This implicit knowledge can be challenging to capture in an automated fashion.

The following message excerpt contains a high density of relevant information about the disease:

*Thanks for writing. The concept of free-floating LAM not being chylous sounds interesting ... haven’t heard of this before and am glad to hear it is not growing. [deleted text] ... event when I get a tx, I will still have the abdominal involvement ... the tx will only relieve the breathing issue. One doctor said a*

*medium-chain triglyceride diet could help because these fats bypass the lymphatic system and are directly absorbed ... but these diets are really yucky and not much info on them ... I prefer to enjoy my meals and have the “belly” as long as I have no other uncomfortable symptoms!*

This passage asserts that:

- Free-floating LAM is not chylous (hypothesis as definition),
- Medium-chain triglycerides are a kind of fat,
- Medium-chain triglyceride fats could help because they are directly absorbed,
- Medium-chain triglyceride diet could help because fats bypass the lymphatic system,
- MCT diets are yucky,
- Taking a prescription will only relieve the breathing issue,
- The diet will help abdominal involvement, and
- Medium-chain triglyceride diet will reduce abdominal distention.

The following excerpt follows up on the issues of chyle <sup>1</sup> is raised above and provides a rich source of definitional and narrative information about chyle origin, production, and side effects.

*My experience with chyle and what I've been told. All people make chyle, it is necessary to have it. Some people have one leak in their lymphatic system that allows it to leak in to the body, some (like me) have several leaks all though-out their lymphatic system. High fat foods really makes it - but all food does make some, some more than others. .... As far as I know, the only thing to do is to go on a liquid diet to shut the chyle manufacture down as much as possible, and just maybe the hole or holes will seal over where lam cells have ate away at the lymphatic system. This has never worked for me, but any surgery I have etc. I do this to better my surgery incision healing etc. Wish I knew the answers*

---

<sup>1</sup>Chyle is a combination of lymph and lipids which in LAM patients can leak into the abdomen due to tissue growths that tear the walls of the lymphatic system.

This passage asserts that:

- Leaks in the lymphatic system enables leakage of chyle into the body,
- High fat food results in more chyle than other foods,
- A liquid diet will shut down chyle manufacture, and
- Reducing chyle will improve surgery incision healing.

As shown above, much of the knowledge in these passages can be broken into two element predicates such as (*isAAB*), (*propertyOfAB*), (*enablesAB*) or (*causesAB*). The identification of a predicate is often dictated by specific word use. Arguments are typically phrases that may be the direct subject or object of a predicate term, but often are not (e.g. surgery incision healing above). The identification of arguments is complicated by the heavy use of anaphors, typically 'it', in patient language.

### 3.2.2 Speech Acts

There are no standard speech act ontologies for online discussion in the literature. Typically, an ontology is chosen that is appropriate to the domain or to the information retrieval task at hand. To better understand how information is conveyed within forums, classification into a 2-level hierarchy of speech acts (Table 3.2.2) was adapted from [KCF<sup>+</sup>06]. This provides insight into the observed range of intent exhibited by users.

Examples from the highest frequency categories are:

- *Social / Life Narrative* “When I went to my usual hospital with shortness of breath, they did not see any pneumo on the cxr. They did a cat scan to look for clots and saw the pneumo. Since I have not had a ct scan in a long time, I am wondering how long I have had a slow leak. My breathing is better than it has been in many, many months. I am taking a duoneb religiously every 4 hours.”
- *Information / Statement* “TS can be passed on, sporadic LAM isn’t. For a disease to be inherited, the genetic mutation needs to be in the germ cells ie the sperm and ovum. The LAM mutation doesn’t show up there so we can’t pass it on.”



General	Nonlabeled Indecipherable Management Welcome Forward Task Quoted Material Signature	Social	Supportive Appreciation Acknowledgment Humor Emotional Material Life narrative
Information	Command Suggestion Hedge Performative Statement Subjective Statement Declarative Rhetorical Question	Questions	Question Wh - Question Y/N Question Reformulation Rhetorical Question
Answer	Clarification Explanation Expansion Correction Narrative	Response	Understanding Check Accept, Yes Answers Partial Accept Partial Reject Reject, No Answers No knowledge answers

Table 3.1: Speech Act Annotation Ontology

- *Questions / Question* “Does anyone have problems with lower back pain? I’ve had this pain for months and the doctor can’t find anything wrong. I had an x-ray and CT back in February and they found nothing. So I just went along with the pain all these months and dealt with it. I went to the doctor yesterday because the pain is excruciating, so again he scheduled me for an ultra sound of the kidney, he thinks it could be the AML, but the blood work he did yesterday showed the kidney function normal. Does anyone have any suggestions?”
- *Answer / Expansion* “No chyle, [NAME], just lots of cysts on my lungs and uncomfortable SOB. Much less since starting Sirolimus. According to the research, as soon as you stop the Sirolimus the LAM cells go back to approx. 85% of what they were prior

to using Sirolimus.”

Life Narratives typically describe a specific experience. Causal knowledge is implicit in the temporal ordering and in rhetorical statements. This information typically presents as existential associations and form the basis for proposing possible causal relationships. In contrast, Statements typically represent the belief of an individual and are more likely to be universal statements conditioned on some context (such as disease state). Statements are much more likely to capture formal clinical knowledge (70% vs. 30%) than patient-specific language or interpretation.

Content in Question e-mails is often qualified with a specific narrative as above. Answers are nearly identical to statements in both content and linguistic structure. The semantic entities referred to, such as symptoms, body parts, interventions, tests are nearly identical across all of these types. However the narrative vs. statement dichotomy of language use has implications for extracting relations among terms as discussed in Section 3.5.3.

There is also only a modest variation in lexical terms between between these contexts. However there is significant difference in the use of tense and aspect between the two clusters of narrative-question and statement-answer. Language use in narratives is typically first person and past or present perfect or perfect progressive tense (e.g. “I’ve had”). Statements are almost always third person and in present progressive tense.

The most significant implication of these properties lies in semantic interpretation; the distinction between instances and generalities is crucial for understanding when and how to interpret the represented knowledge.

A visual summary of the distribution of speech acts is illustrated in Figure 3.2.2. The color of each square encodes the density of medically-relevant content within each speech act category.

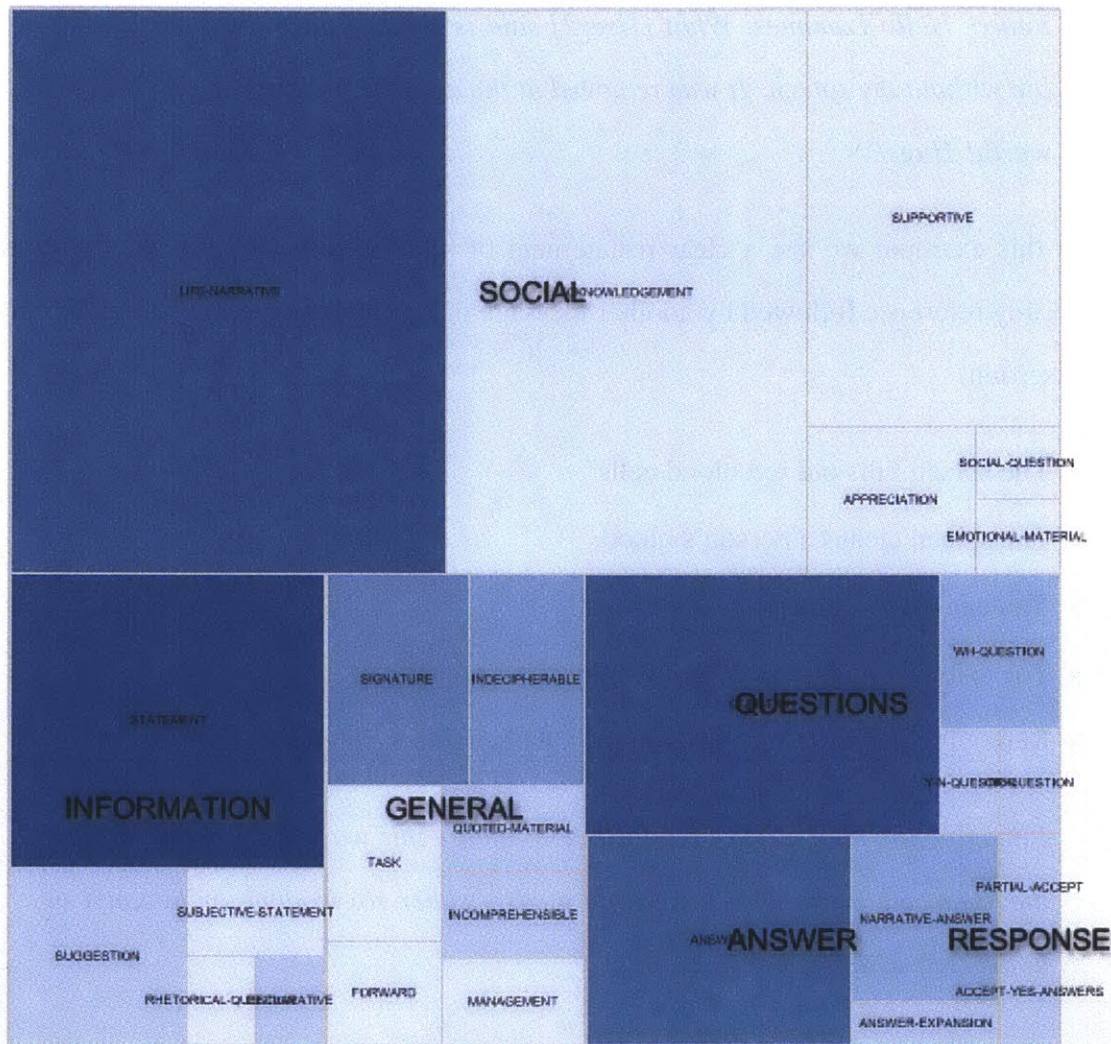


Figure 3-1: Density of medical content by speech act

### 3.2.3 Anaphoric References and Implicit Context

A challenge that occurs often in this corpus, like many, is the frequency with which anaphoric references separate related concepts across sentence boundaries. It is rare that a single parse tree with a head verb is alone sufficient to identify the knowledge being conveyed.

*The spleen is the thing that kills old red blood cells and cleans your blood. You can live without it as what it does can be done by other organs, but doing that makes you more vulnerable to infections. [url-ref] love [User2].*

*Reply: Hello Lammies, What [User2] said is right. I have lived most of my life without my spleen. It was removed at the age of 8, for me. Just my 2 cents worth! Hugs!*

In this example we see a clear restatement of scientifically valid knowledge with a supporting reference followed by another message that provides experiential validation for the assertion.

- The spleen kills old red blood cells
- The spleen cleans a person's blood,
- You can live without a spleen
- The spleen's function can be done by other organs
- Removing the spleen makes you more vulnerable to infection

*I imagine the pain might be a little less intense this way too. I'm still hoping they send me home with good pain meds - either roxycodone or fentanyl or both. That will knock it right out.*

A richer form of anaphoric reference comes in the form of an assumed context or semantic understanding. Here a person can easily infer that roxycodone (oxycodone) or fentanyl are pain meds and that they will stop the pain, but binding these terms to *they* or *pain meds* is ambiguous without additional knowledge.

### **3.2.4 Identifying Clinical Terminology**

Patients, particularly those with rare diseases, are often highly educated and have a sophisticated clinical knowledge of their disease(s). It stands to reason that a many terms in a patient forum could be mapped to a standard medical terminology.

The Berkeley Bioinformatics Open Source project [SAR<sup>+</sup>07] contains an simple ontology of over 800 clinical symptoms. For the manually labeled content, less than 10% of the

total contained a term in common with a symptom from the ontology. Over 30% did match to a keyword from this corpus.

The UMLS Metathesaurus [Ano09] produced by the National Library of Medicine contains over 200,000 Preventive or Therapeutic procedures, 2,600 Signs or Symptoms, and over 10,000 disease conditions. Of these, 510 directly match with the corpus. This accounts for less than 30% of the labeled terms.

Many of the terms that do not match in the UMLS perform a nearly identical semantic function in the discourse. For example, “lung collapsed”, “collapsed lung” are both common terms for the UMLS concept “pneumothorax”. “feeling like there is a weight on your chest” is a commonly expressed symptom may match to one or more UMLS concepts including “pleurodisic pain”, “shortness of breath”, or “pneumothorax”. This parallel use of language has been remarked in other work, such as Lieberman and Moore [ML09].

The OpenCalais<sup>2</sup> semantic tagging system also contains a simple medical ontology and can be used to tag medically relevant terms in open text. Such systems are often a good source of a-priori knowledge about the content of a corpus. However, OpenCalais provided a surprisingly low recall (less than 2% of the terms) albeit with a respectable precision of 85%<sup>3</sup>.

### **3.2.5 Topic Models**

General spelling correction methods (e.g. aspell) distort many of the domain-specific terms and abbreviations of interest found in the corpus, therefore a mechanism for grouping known terms with unknown abbreviations and misspellings will be helpful in classifying unmatched terms in the text.

Generative topic models [SG07], such as Latent Dirchelet Allocation (LDA) [BNJ03], are unsupervised statistical models of corpus content. Topic models represent a corpus

---

<sup>2</sup><http://opencalais.com/>

<sup>3</sup>Given the poor initial results, a more detailed examination of OpenCalais was not performed.

Condition	Symptom
Disease	Breathing
Cancer	Quality of Life
Reproduction	Bone Health
Infectious Disease	Chyle
LAM	Blood Stats
Environment	Digestive Symptoms
Treatment	Breathing Symptoms
Sleep	Pulmonary Symptoms
Tracking	Weight
Surgery General	Body Pain
Clinical Trials	Lung Symptoms (x2)
Sirolimus Trials	Mental Conditions
Respiratory Meds	Other
Asthma	Oxygen & Finance
Physical Therapy	Oxygen Equipment
Food	Oxygen Accessories
Diet	Oxygen and Flying
Tests	Transplant Logistics
	Physicians

Table 3.2: 36 Medically Significant Topics

as a bag of terms drawn from one or more topics. Each document is associated with a distribution over topics, and each topic a unique distribution over terms. Conceptually topic models are an extension of techniques, such as pLSA.

The Mallet machine learning library [McC02] was used to train an LDA topic model of the top 200 concepts in the LAM listserv corpus. Thirty-six (36) of the topics contained highly-relevant clusters of unigrams that reflect common community topics related to the disease and/or possible interventions. These topics were manually labeled, often using the highest weighted word in a given topic’s word distribution. These labels are summarized in Table 3.2.5.

Off-topic categories primarily included common topics for social discourse (e.g. “dogs”, “weather”, “lamposium”), high-frequency posters<sup>4</sup>, and filler topics.

<sup>4</sup>One interesting observation was that many of the top posters had signatures associated with a specific topic, e.g. [name] [hometown] [friend] [husband’s name] lam boys husband time prayers [name] [lastname] oregon trial forward hear family today hong university seattle patients touch find nice night kids making good great works boy advice year turns makes breathing problem

For example, the top terms for the category Chyle includes <sup>5</sup>:

chyle fluid fat lam lung abdomen system  
pleural octreotide lungs diet problem left  
lymph nih effusion lymphatic low drained  
pleurodesis surgery day chest years issues  
belly body free feel side drain experience

As seen above, topic terms are a mix of semantic types such as body parts (lung and abdomen), substances (lymph and chyle), interventions (diet and pleurodesis) and other modifiers (daily and left) which are associated with chyle, but are not medical concepts in their own right. Topic models providing a strong prior over the likelihood that an observed term in the corpus is from the same semantic topic as another term in the corpus.

Recent work on topic models has developed algorithms that capture sequential dependencies in the data, such as N-gram Topic Models [WM05] and HMM-LDA [GSBT04]. N-gram topic models extract both unigrams and n-grams associated with that topic. Initial analysis of this algorithm on the corpus yielded unsatisfactory results, most likely due to the modest size of the corpus and the large variation in contextual term use. HMM-LDA was not applied, but is also likely to suffer similarly.

### 3.2.6 Corpus Summary

This analysis of the corpus identifies four general characteristics of the data that present both challenge and opportunity for automatic identification and extraction of knowledge.

- *Topic Diversity* - Many of the messages are social in nature and carry little information of relevance to the disease itself. While speech acts (i.e. Life-Narrative, Answer) can provide a positive or negative prior, they are not significantly discriminative of relevant content.

---

<sup>5</sup>Patient names heavily associated with discussions of chyle were in this topic, but removed from the listing.

- *Discourse Structure* - The speech acts are equally split between narrative and declarative modes. As discussed, narrative knowledge is an assertion of an observed temporal co-occurrence in daily life. Causality is implied, but rarely declared. Declarative knowledge, by contrast, are overt statements of an individual's belief about universal association or causation (perhaps conditional on a specific circumstance). Verb tenses are a cue to the type of communication that is taking place.
- *Long-distance Dependencies* - A sentence or paragraph often express a set of inter-related concepts. Messages can encode binary semantic relationships, but they are rarely cleanly expressed within the context of a given phrase or sentence.
- *Linguistic Impairments* - A more prosaic challenge of this corpus, and forums or e-mail in general, are the varied impairments in syntax, spelling, and the common use of abbreviations which confound extraction techniques based on parse trees, word choices, etc.

The following section identifies a target representation for the knowledge in this corpus and provides examples of manually derived instances of this formalism. The section following will outline a strategy for automatic population of this representation, including preliminary results from a system implementing the strategy.

### **3.3 Representation**

While speech act analysis provides some insight into how patient medical concepts are communicated, it does not address the representation of those concepts. The selection of a semantic representation in any context is highly dependent on the inferences we wish to make with it. The representation chosen must strike a balance between richness of human conceptual representation and tractable inference. The goal for the representation developed here is an aggregate model of patient theories about disease etiology. Ideally



this representation and the acquired data can, with some human intervention, enable litmus testing against observational data and simple counterfactual (“if-do-then”) reasoning in support of patient community dialog and individual decision making.

A vast array of representations exist to capture aspects of human mental models, three major approaches formal logic, frame-based, and relational networks.

Attributed grammars that extract logical sentences from syntactic parse trees have been highly successful in limited domains. The logical formalisms, such as the Event Calculus [KS86], have been used successfully to represent a broad array of knowledge about daily life found in narrative discourse [Mue06]. Typically, subtle issues of human representation are difficult to represent without significant machinery and the required precision of the formalism make it a poor match for intuitive representations. Moreover, the requisite deep parsing and entity extraction required to populate logical representations make them an impractical tool for this corpus.

Frame semantics [CNP02] is a less precise representation designed to capture essential characteristics of the conceptual structures invoked by “frame-bearing” lexical heads. These conceptual structures identify the expected types of phrase-level constituents that are brought into relation by that head term. (e.g. “the boy threw the ball to the dog” => [*THROW* Agent: boy, Theme: ball, Path: to the dog]).

Frames are excellent representation of the rich structure of situations, events, objects and their properties. Frames afford predictions in terms of the type of expected role fillers and restricting possible relationships between frames in a text. Some success has been achieved [SM05] in automatic parsing of text into this representation. Unfortunately, semantic parsing into frame representations require annotated corpora and existing corpora for training such systems is limited both in size and coverage and annotation is extremely expensive.

### 3.3.1 Semantic Relations

All of these representational modalities contain a common notion of a relation between concepts (e.g. (*CapableOf**throw*)). Lexical features and lexical semantics such as hyponym (kind of) and hyponym (part of) have been detailed in resources such as WordNet and VerbNet which can provide constraints on possible semantic relationships among lexical terms. However, knowledge representation is typically concerned with relationships such as *cause*.

Constituents of a relation are typically organized in the form of relational ontologies that represent the structure of concepts into categories such as treatments, drugs, symptoms, etc. Semantic category recognition is a branch of information extraction that seeks to label n-grams in open text with their semantic type. These types then constrain the likelihood of a given relation. The methodology for extracting primitive relations from the corpus are discussed in the next chapter.

### 3.3.2 Networks of Relations

Theories of diseases can be captured by networks of relations that represent causal sequences, alternative explanations, related evidence, etc. The most exhaustively explored representation for networks of relations are Bayes Nets [PS88], a graphical model. Bayes Nets represent a single relation type, *association*, and augment that with parameters for probability distributions that represent the nature of the association.

The complex structure of the disease domain may perhaps more suited to richer probabilistic representations such as Probabilistic Relational Models [Kol99] or Markov Logic Nets [RD06]. However the learning and inference procedures for these representations are expensive and complex and, as described above, are likely to require too much accuracy in detail than is accessible in the corpus via current information extraction techniques.

The graph structure of a Bayes Net is a compact encoding of the joint probability over a set of variables. A link between the nodes represent a conditional dependence, and the lack

implies conditional independence. Nodes may be measurable quantities, or abstract hidden variables. Choosing the nodes that make up the graph is part of the black art of using these models. This formalism affords inferences regarding the predicted value of some variables given data about the others. Independence testing is one way to test the validity of a given graph against real-world data.

The conditional independence assumptions in a Bayes Net are relatively for ordinary people to easy to understand, although even experts fail to exhibit accurate intuitions about the implications of a causal network.

Bayes Nets alone do not represent the kind of causal knowledge that would enable the desired counterfactual reasoning. Developments over the last two decades have led to extensions of the Bayes Net formalism which give the directed links between two vertices in the graph a causal interpretation [SGS00] [Pea00] causal models have been highly influential in a variety of disciplines most notably in fields such as econometrics and the health sciences. Causal Bayes Nets has been adopted by the cognitive science community as a formalism for characterizing human causal modeling and learning [Gly03].

Approaches to causal learning fall into two basic methodologies. The first is constraint based, exploiting observed conditional independencies to restrict the universe of plausible graphs, or causal explanations, of the observed systems. The other methodology draws on Bayes learning of the parameters that govern the value of variables in the graph. Causal Bayes Nets then afford inference that computes the causal effect of interventions on the graph. Without experimental controls, however, all causal learning and inference requires exogenous assumptions.

- Causal Markov Condition - a variable  $X$  is independent of all other variables conditioned on its parents.
- Faithfulness - The graph represents exactly the independencies of the underlying system.

- Causal Sufficiency - All the common causes of a measured variables have been measured.

Given these assumptions, a variety of algorithms can be applied to learn the parameters of the graph and the direction of causal influence. All of these algorithms require that the variables to be represented are selected and that constraints on possible graph structures be provided (or learned via independence tests of the data). This pre-supposes a high degree of insight into the domain in question. Thus, the accuracy of the experimenters intuitive model of the domain is a key factor in the efficiency and effectiveness of the subsequent research.

Given an accurate causal graphical model with or without parameters, it is then possible to perform inferences using the graph to suggest new measurements, experiments or hypotheses about the disease. Anecdotes from the LAM community suggest that patient self-report about some phenomenon (e.g. shortness of breath is correlated over the short term with hormone fluctuations) can often precede clinical or laboratory discovery of the underlying phenomenon. Patient observations and the available of widespread electronic medical records or longitudinal, observational platforms [FMWH08] can potentially make these anecdotes something that can be acted on directly by researchers; accelerating the development of insights into complex disease processes.

### **3.3.3 Target Representation**

To balance representation coverage and tractable inference, a subset of the knowledge outlined in 3.2.1 is addressed by the following representation:

- A Causal Graphical Model formalism
- No commitment to modeling probability distributions, but to acquire networks of associations and causal relations that form the basis of constraints over plausible causal structures

- A simple ontology for constituent types: condition, symptom, and intervention. A hidden variable condition, for instance, may be “vitamin B-12 deficiency” which causes the measurable symptom “low B-12 levels in blood serum”.

The goal of information extraction from the LAM corpus is to identify valid constituent concepts from the domain and hypothesize the associational or causal relation between them.

### 3.4 Information Extraction

In Section 3.2.6 four key challenges related to extracting frame-level knowledge from the corpus were introduced: topic diversity, discourse structure, long-distance dependencies and linguistic impairments. The goal of this section is to demonstrate a procedure for mapping from this corpus to the representation introduced above. We subject this procedure to the additional constraint that it minimize the expensive hand annotation required by the majority of information extraction techniques.

As discussed above, the dominant source of information for extraction is a combination of n-gram semantics (domain-relevant, semantic type), lightweight syntax, and a-priori knowledge that can be used independent of accurate parse trees. In health forums we can take advantage of the vast amount of work in formal medical ontologies and the nature of the forum should provide an a-priori bias towards the inclusion of terms from the medical ontology.

The approach here draws from four areas of information extraction research:

- Topic modeling – Topic models compensate for long distance dependencies and a lack of contextual diversity, topic models are employed to unify constituents across all messages into a specific topic.
- Conditional Random Fields (CRF) – CRFs [LMP01] are a well established probabilistic model that have had great success in sequential labeling tasks such as POS tagging or

named entity extraction. Here they serve to label the specific semantic type using local linguistic content to separate the different types of terms in a specific topic category.

- **Leveraging External Knowledge Sources** – A great deal of work use general hand-built resources like WordNet or the UMLS to boost the performance of purely statistical techniques. Knowledge is a powerful prior for these learning algorithms.
- **Bootstrapping** – Bootstrapping [Jon05] has been explored in the domain of probabilistic models as well as in template-based information extraction as a way to use unlabeled data to boost the discriminative power of a small set of positively labeled data. This allows us to generalize from a small seed set to an entire dataset.

This approach requires three sequential operations on the corpus:

1. **Topic Modelling** - A topic model identifies the terms within sentences that are deemed to be generated by a given model.
2. **Constituent Extraction** - Relational constituents labeled using the UMLS are used as supervised examples to train a Conditional Random Field as condition, symptom, and treatment.
3. **Relation Classification** - The context (one to three sentences) in which pairs of constituents have been identified are classified using Naive Bayes. The supervised training set is formed from UMLS semantic relations and due to its sparse coverage, bootstrapping is used to boost the performance of the learning algorithm .

The Langutils toolkit [EL05] provided tokenization, POS tagging, and phrase chunking (shallow parsing). Sentence boundary identification uses simple regular expressions. Topic model labels were assigned using the tools and methodology described in section 3.2.5 including n-gram and hierarchical topic model experiments.. The CRF++ library was used for CRF training and classification.

### 3.4.1 Extraction Example

Revisiting a prior example will serve to ground the nature of the challenge, particularly for a significant recall of the actual relations expressed in the text.

*The spleen is the thing that kills old red blood cells and cleans your blood. You can live without it as what it does can be done by other organs, but doing that makes you more vulnerable to infections. [url-ref] love [User2]*

*Reply: Hello Lammies, What [User2] said is right. I have lived most of my life without my spleen. It was removed at the age of 8, for me. Just my 2 cents worth! Hugs!*

In this example from section 3.2.3 we identified a set of simple factual sentences. They can be reframed in terms of the primitive relations of the representation.

- “spleen” causes “old red blood cell death”
- “spleen” causes “clean blood”
- “mortality” independent-of “exists spleen”
- “clean blood” independent-of “exists spleen”
- “old red blood cell death” independent-of “exists spleen”
- “remove spleen” causes “vulnerability to infection”

These fragments yield the graphical structure illustrated in figure 3.4.1. The graphical model includes the direction and nature of the influence assuming the variables are all cast as booleans. For example the existence of a spleen positively influences clean blood; its absence makes you vulnerable to infection. The surgery and infection nodes are implied, but not explicit in the text. One challenge to the formalism is that variables like alive are exogenous in this graph, but are in fact highly influenced by variables like infection. The structure of the graph is highly dependent on how we draw the boundaries of the system

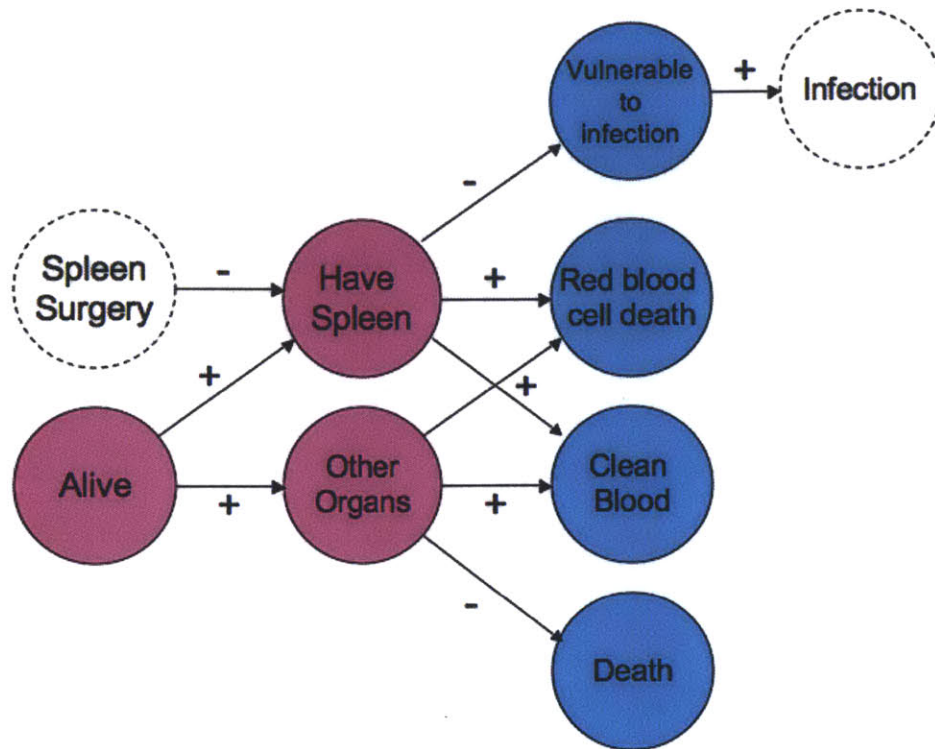


Figure 3-2: Fragment of a Causal Model

and thus presents challenges for aggregating patient theories that are based on different conceptualizations of the same set of phenomena.

Generating this list required restructuring the actual n-grams in question as well as interpreting the language of the phrases. One interesting property is the conflict between the spleen causing and being independent of cleaning the blood. This is evidence of multi-causality as illustrated in the model fragment

What is actually feasible to extract from these phrases? Clear mappings between linguistic phrases and medical concepts from the UMLS include “spleen”, “red blood cells”, “cleans your blood”, “live”, “organs”, “vulnerable to infection”. The second passage includes “lived” and “spleen”. In this case false positives are highly likely due to insufficient context attached to nodes, For example “spleen” *causes* “red blood cells” vs. *causes* “kill red blood cell death”. Having a clear semantic mapping from verbs to relations would



boost performance, for example “spleen” *kills* “old red blood cells” => help here as  $(killXY) => (causesX(deathY))$ .

The linear structure (*condition* “live”) most of my life without my (*symptom* “spleen”) would not yield a causal connection but without is an existential example that the condition and symptom are not causally related and are in fact likely independent.

### 3.4.2 Constituent Extraction

The problem domain of constituent extraction is identical to that of named entity recognition or semantic category labeling. The aim of these techniques is to train a model that can identify an appropriate labeling of n-grams in the text.

Training labels were identified by matching instances of semantic types from the UMLS database to n-grams in the text. Because the labeling is not complete, the system is trained only on messages with positive instances and that model is then applied to the unlabeled data. The goal was to over-generate candidate instances that are be filtered during relation classification so as to pick up terms that are misspellings, abbreviations or patient language outside the UMLS. For example a common shorthand for “lung transplant” was “lung tx” or just “tx”. Results of training and applying the classifier are discussed in section 3.5.

A Conditional Random Field model that models the probability distribution over labels conditioned on an observation sequence. The distribution is computed from a set of feature functions defined over local information including lexical terms and part-of-speech.

Initial experiments with this mode resulted in an overwhelming numbers of false positives. The classifier was significantly under-fitting the categories of interest. It effectively was classifying a more general semantic category. For example terms were generally about state of being vs. medical treatments or about symptoms of their dog rather than their own condition. The relevance of a term or phrase to the domain is highly dependent on non-local context. Non-local context means the topic of the message and whether that dictates the affinity of a term for a medically relevant topic rather than a social one. The topic models

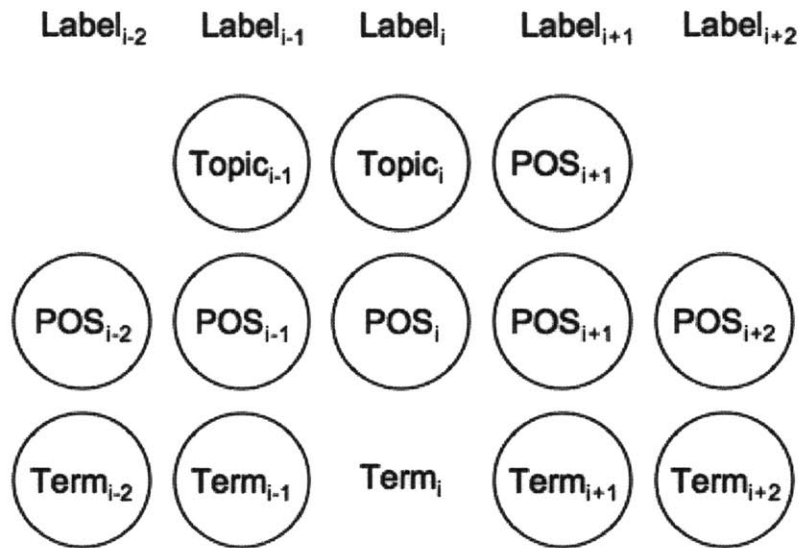


Figure 3-3: Features for Conditional Random Field

discussed in section ?? provide exactly this constraint.

The addition of topic models however, lead the model to over fit to the exact lexical terms contained in the UMLS database. By removing the feature function that depends on the current lexical term, the resulting model is more balanced, but still over-fits the large training set.

The final model of Figure 3.4.3) integrates the topic information into the CRF classifier to merge local and long-distance constraints while removing direct dependence of a label on its lexical term.

While not illustrated, two and three-gram sequences of these three data types were included in the feature function template as well as label bigrams ( $[-2,-1]$  and  $[-1,0]$ ).

### 3.4.3 Relation Identification

Once constituent n-grams and their types have been identified, the relationship between them must be determined. Approaches for relational extraction have traditionally depended

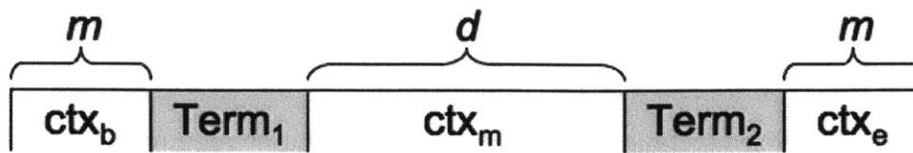


Figure 3-4: Relation Context Window

on pre-determined context such as subject-verb-object syntactic structures, leveraging semantics of the verb to determine the relationship between subject and object (trivially, “pleurodesis causes horrible chest pain.”). Generalized pattern-based approaches [GM02] [ECDK04] have met with great success for simple relations clearly expressed within a given phrase or sentence and commonly expressed in web content. Initial results indicated that the variation in the local context of this corpus relative to the corpus size limits the effectiveness of the pattern based approach.

A bag of words approach is more robust to variance in the lexical context of two terms. Naive Bayes classifiers have demonstrated remarkably robust performance on similar tasks. The features used to train the classifier include the windows of terms that capture the lexical context of the n-gram pairs, illustrated in figure 3.4.3. Additional features include the types of the two arguments, and their lexical order.

Regions of text with extract n-gram pairs are allowed to extend over sentences, but the total size of the window is upper-bounded. The windows are parameterized by:

- The width  $w$  of the term sequence between the n-grams,
- The start  $s$  and end  $e$  term context of size  $m$ , and
- The constituent types and order.

If the location of a word relative to the pair is highly discriminative, the model will fail to take advantage of this information. Variations in feature set might include tagging each term with its location ( $term|loc$  for loc in left, middle and right) or remove the n-gram terms from the feature set to avoid over-fitting to highly common terms like “oxygen”. The results presented include just the untagged lexical terms including the n-gram terms.

### 3.4.4 Bootstrapping

Naive Bayes is a supervised technique, and the manual annotation required to provide sufficient coverage of the corpus for semi-supervised learning techniques is exceedingly expensive. To minimize the cost of generating a training set, and to identify a general strategy for other health forums, hand-build external knowledge bases can be used to generate positive examples of constituents and relations as discussed in [KMB<sup>+</sup>05]. In this experiment the UMLS was the source of both typed n-grams and the causal relations between them.

However, discriminative classifiers like CRF and Naive Bayes perform poorly when given only positive examples, typically by over-fitting. One solution to this problem is leveraging unlabeled data to generate negative examples. This bootstrapping technique has been applied both to pattern-based approaches and to discriminative models.

Bootstrapping techniques are trying to find the maximum likelihood estimate of a discriminator between one or more classes given a small set of positive examples of that class. In this case, the presence or non-presence of one or more relations. The schemata for bootstrapping includes training on a seed set, applying the classifier to the unlabeled corpora, evaluating which of the resulting labels maximizes some function and adding those instances back to the seed set [Jon05].

In pattern-based approaches to classification, the new pattern which retrieves the greatest number of seed instances is added to the pattern seed set. The most frequent extracted instances not in the seed set is added to the instance set. Iteration continues to convergence within some epsilon change in the set of labels produced.

Another common approach is to use expectation maximization over unlabeled corpora [JNRM99] [YYP08]. Expectation maximization is used to estimate the max likelihood parameters of a model in the presence of hidden variables, in this case unlabeled data. The implementation of EM here is very similar to that described by Yetisgen and Yildiz.

## 3.5 Preliminary Extraction Results

The application of the above techniques were evaluated both against the initial manual tagging of the dataset as well as manual evaluation of the final results of each stage of analysis.

### 3.5.1 Constituent Identification

The UMLS database of symptoms, treatments, and conditions was compared to the corpus to identify 24,372 instances of 794 medical terms. These are used as labels for the training set.

Constituent identification consists of three steps:

1. Generating a training set using only sentences containing known labels,
2. Training a CRF model on this training set ( $C=0.5$ ), and
3. Label the entire corpus using the resulting model.

The classifier re-captured 50% of the original UMLS labels and 305 new instances of 131 novel concepts. A random example of new terms from all type classes include:

"kidney cyst" "low fat" "awareness of"

"counter" "catheterization ,"

"asthma" "lobe transplant ," "mouth"

"physical therapy" "round type" "illness"

"ischemic heart" "outage" "information"

"very low fat diet" "heart lung"

"lung transplant surgeon" "drainage therapy"

A manual classification of the identified terms (including the UMLS matches) indicates a precision of around 75% overall as shown in figure 3.5.1. The general term category consists primarily of fragments of phrases.

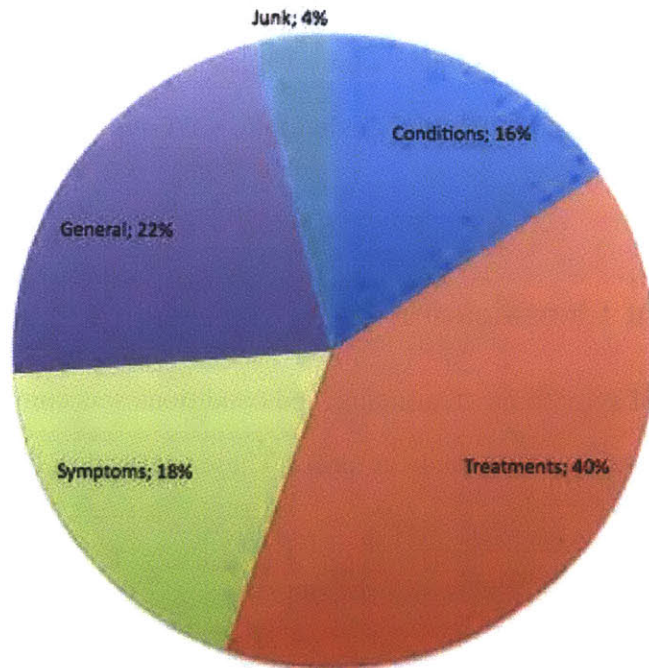


Figure 3-5: Constituent Precision

Few of the newly labeled terms are highly frequent in the corpus overall. The corpus word rank of the non-stopwords found is partially illustrated below:

```

("lung" 97) ("call" 117) ("start" 132)
("care" 201) ("raise" 206) ("pain" 225)
("transplant" 249) ("hospital" 254)
("heart" 322) ("chest" 337) ("disease" 342)
.....
("yoghurt" 7274) ("marrow" 7426)
("investigation" 7791) ("pancreas" 8300)
("low-fat" 9393) ("coronary" 9515)
("tranplant" 10129) ("dispute" 16250)
("catheterization" 17774) ("outage" 18187)
("cantaloupe" 21842) ("ischemic" 61271)

```

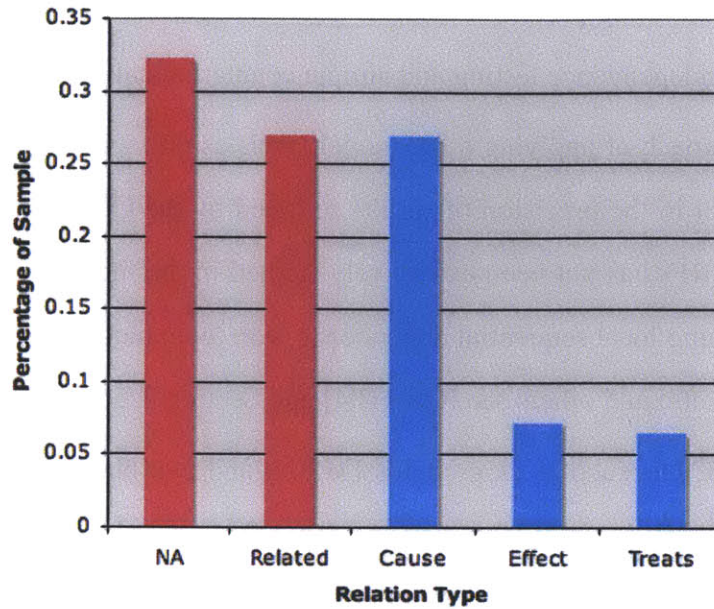


Figure 3-6: Relation Labeling

Without a complete training set, it is difficult to evaluate the recall of concepts over the corpus. The manually labeled data used to characterize the knowledge in section 3.2.1 serves as a small sample. Of the manually identified semantic types, less than 10% were identified by the classifier.

### 3.5.2 Relation Extraction

The quality of relation extraction is highly dependent on the recall of constituent labeling. Low recall reduces the probability of two constituents being identified within a text window that expresses a relation between them.

### 3.5.3 Discussion

The techniques described above extract a skeleton representation of patient theories embedded in a noisy discussion forum. It avoids reliance on expensive hand-annotated datasets through a process of bootstrapping, unsupervised learning, and leveraging external knowl-

edge sources. The resulting precision result is good enough that it can form a reasonable baseline for basic independence testing and simple counterfactual hypotheses.

The hybrid approach of applying topic model labels as priors in CRF training leads to a 50% improvement in the precision of results returned by the CRF. The use of topics as explicit priors in CRFs has not been extensively studied in the literature, however unified approaches combining local sequential dependence with topic models has been the subject of work by Tenenbaum and others see [GSBT04] and Section 3.2.5.

Prior work on CRFs such as skip-chains [MS04] and two-phase CRFs [KM06] have attempted to address the issue of long distance dependencies with some success. The fundamental challenge in these models is that long distance context must be significantly reduced to avoid an exponential blow-up in the model size. Systematic assumptions about long distance dependencies are in fact the central design problem for accurate and tractable extraction from corpora like the LAM listserv.

In this approach, the content of adjectives and prepositional phrases which qualify the subtype or domain of validity for a given root n-gram is poorly represented. Combining these n-grams to unify relations is thus prone to overgeneralization. Techniques such as parse dependency mapping can be applied to add contextual qualifications to the extracted n-grams.

The relation classification suffers from extremely poor recall; in the labeled training set only approximately 10% of the manually labeled relations were identified. Several factors lead to this poor result:

- The UMLS seed set provides insufficient coverage to enable proper generalization over the distribution of actual contexts.
- Constituent identification - the CRF model's ability to generalize over constituents not in the original UMLS dataset remain limited.
- Anaphoric references - no resolution was performed. Forming lexical windows by doing a best-guess replacement of anaphoric references with referent terms would likely



increase recall; the impact on precision is not clear.

While a substantial amount of information is accessible using these techniques, a much larger body of knowledge remains locked away. Future work along these lines will require a more aggressive cross validation of held-out data, multiple annotators to establish human agreement levels to properly characterize the preliminary analysis here.

The most promising avenues for improving on these initial results include the following strategies.

1. Leveraging a richer body of semantic knowledge about word types (e.g. from Word-Net) to boost constituent extraction.
2. Deeper parsing of the content (e.g. link-parser, parse-dependency mapping) to complement the lightweight statistical techniques. Better identification of the context of an assertion will help to avoid overgeneralization but also to generate priors over semantic class probabilities.
3. More aggressive investment in hand-labeled seed data to boost the performance of the semi-supervised techniques used here.
4. Directly engage patient communities. Many web sites have been launched in the last several years which capture compact representations of symptoms, treatment terms, etc. Using patient as sources of supervised feedback and filtering may be a source of significant improvement in constituent recall as well as relation identification.
5. Exploit narrative vs. declarative style. Verb tense and aspect anecdotally are excellent features for discriminating the two styles of text; this context should substantially bias the prior probability that the text order of concept pairs represent the direction of the relation.
6. Explore the suitability of more advanced representations such as probabilistic relational models, to enable reasoning over other relation types such as part-of or is-a to

better capture the conditional dependencies of the underlying causal hypotheses.

### 3.6 Implications

The difficulty of extraction actionable knowledge compounds with the difficulty of grounding each of these narratives. Even if an ideal system could identify prevalence of successes and failures with specific therapies for specific conditions or symptoms (e.g. [?]), such a system would suffer from the following biases.

- **Selection Bias** - Evidence suggests that less than 10% of all patients participating in an online forum report their experience back to others, a substantial and likely unpredictable selection bias.
- **Confirmation Bias** - Individuals form strong opinions, often from limited data, and seek information that validates a preexisting opinion. This appears to be a common factor in some, but not all, forum discourse.
- **Accuracy** - For the reports we have, we do not know the quality of the reported outcome and thus it's veracity as a datapoint for an aggregated summary is suspect.
- **Lack of Stratification** - Most forums do not have, or make available, background information on participants. Thus if a factor is determinative in identifying the subpopulation that responds well, or has a side effect, that information is not available.

To combat these biases and create knowledge that is more directly actionable, we require a better sample of the population, decreased confirmation bias, and improved fidelity over treatment-outcome associations. Empirical outcomes from controlled experiments are the ideal mechanism for probing the tantalizing bits of causal knowledge embedded within health social media discourse.

## Chapter 4

# Aggregating Patient Reported Outcomes

One of the secondary goals of Aggregated Self-Experiments is enabling patient communities to accumulate evidence that documents a patient anecdote to provider or clinical research communities. Quantified datasets can clearly characterize an anecdote in a form that might convince the provider or researcher to change practice, behavior, or direction.

LAMsight is an early web-based system I built as part of my action research with the healthcare system. I developed the tool to support a patient advocacy organization, the LAM Treatment Alliance and their patient community of women with the rare disease Lymphangioleiomyomatosis (LAM). The long term goal was to enable community participation in the research agenda through the volunteering of background survey data and longitudinal "data journals" that documented the observational evolution of symptoms over time (similar to [FOV<sup>+</sup>11]). The site also facilitated community-researcher community as described in the next section.

This chapter describes LAMsight as well as the development, execution and results of a research study initiated by patients who used the tool to generate a novel population-level dataset that successfully advocated on their behalf to the research community. This case study is an early, yet compelling example of one way patient communities can leverage emerging tools to augment anecdotal observations using both large numbers of patients

and concrete data to change the way biomedical research is pursued.

Personal Experiments represents an evolution of this strategy, but emphasizes a process of individual self-improvement instead of the collective “data volunteerism” of LAMsight. Therefore this chapter should serve to enrich the context of the subsequent work, and as a motivational case study for those readers interested in this use of patient-reported outcome data, but is not necessary to understanding the primary contribution of the dissertation detailed in Part II and Part III.

## **4.1 Origin of LAMsight**

Starting in 2007 I interviewed patients, clinicians, and researchers from the LAM community in collaboration with the LAM Treatment Alliance to better understand the clinical research process. We were looking to improve our understanding of the process of innovation and discovery and to identify places where patients could apply leverage to accelerate research around the disease.

The spark point was the observation that advocates felt that they and other patients often noticed something about the disease ahead of the clinical researchers, but it was ignored until it happened to show up as a signal in one or another piece of clinical research. The challenge was that when data was recorded, it was often not the primary goal of a study so it was only accessible to the investigator who ran the study and maintained the patient registry. If they chose not to report it, no other investigator could benefit from that patient data.

Out of this interview process, a project called LAMsight was conceived to address the following goals:

- Create a global index of all LAM patients
- Collect a minimal amount of data to help identify patients for research studies
- Enable and encourage patients to ‘break down silos’ by sharing information on the site



Figure 4-1: The LAMsight Home Page

- Enable patients and researchers to generate their own surveys to "ground narrative" with data
- Support patient formation and test of hypotheses using the collected data

## 4.2 Design and Features

LAMsight provided several features highlighted in the following figures:

- Collect. A facility for filling out and reviewing surveys and diaries. Diaries were repeated surveys. (Figure 4-2)
- Explore. A tool for simply constructing queries against multiple user-generated surveys, creating sub-populations based on their answers to specific sets of questions, and comparing the two. (Figure 4-3)

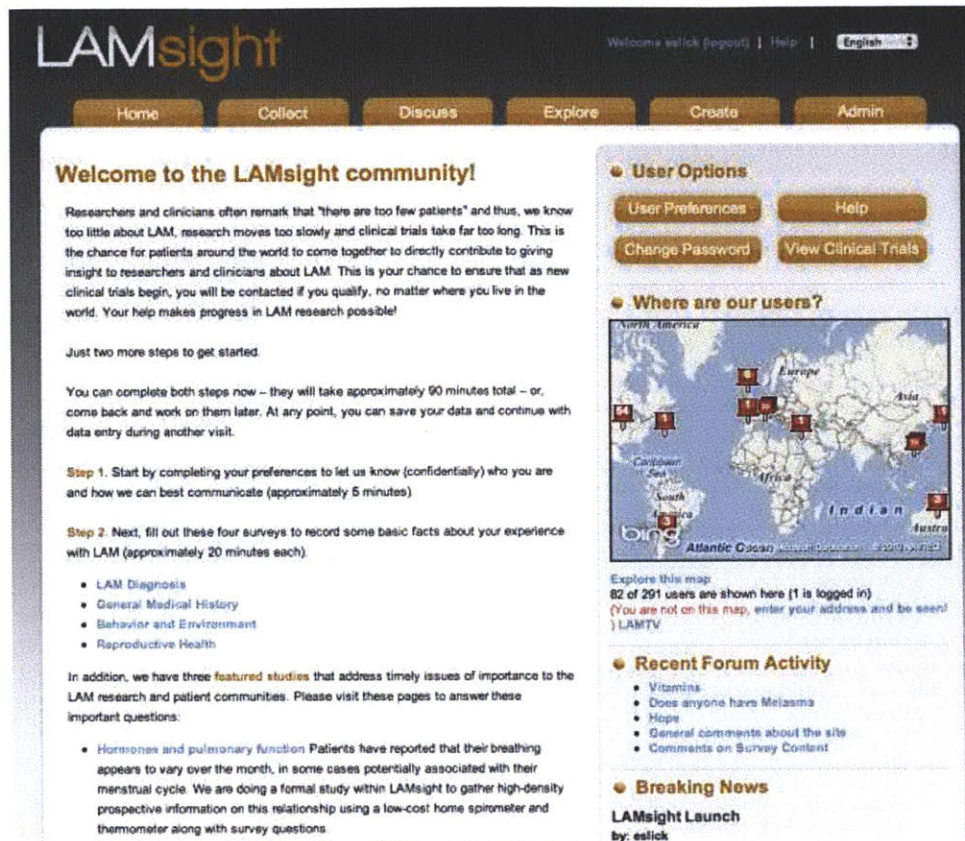


Figure 4-2: LAMsight Data Collection

- Create. An editor open to patients to generate their own surveys. (Figure 4-4)
- Forum. Users were confused by the use of an external Google forum, so we provided a simple forum facility built into the site.
- Settings. A rich set of privacy and context options for notifications was provided.

LAMsight also attempted to implement some novel features. We noted that many questionnaires contained duplicate questions. The architecture of LAMsight allowed assembling topic-specific surveys as sets of questions groups which could be shared; if you filled out a group in one background survey, it would be pre-populated in all the others. LAMsight's database maintained a complete provenance of all the data, including a change history for every answer.

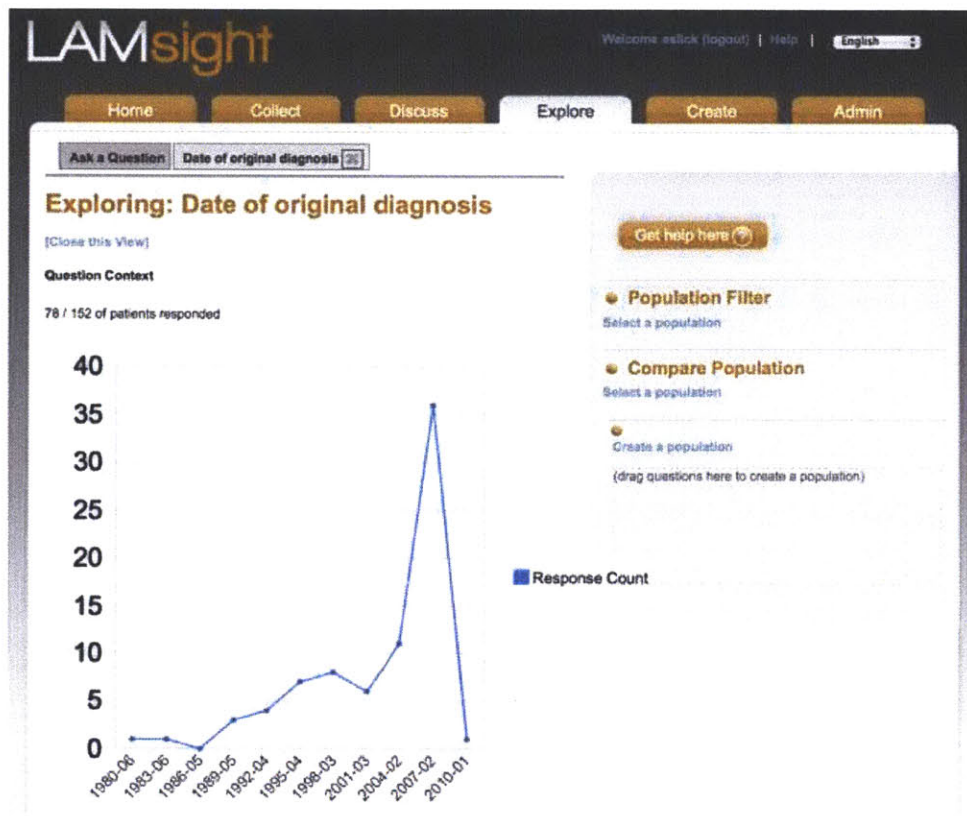


Figure 4-3: LAMsight Data Explorer

### 4.3 Community Reception

The reception of the initial concept when presented at patient meetings or on patient mailing lists was extremely positive. Advocates from country or organization-specific lists around the world, including the US Tuberous Sclerosis Alliance, volunteered to promote LAMsight content on their lists. Many patients expressed that there were experiences they wanted to share with one other, and to document for clinicians, that were difficult to do with existing tools and processes.

Clinicians in the community were initially highly skeptical of the value of patient data or a platform. Several conveyed in personal discussions, that "we already know everything because we all know each other and share all the information as it comes out."<sup>1</sup> Another researcher said that the patient-provided data didn't matter because they couldn't publish

<sup>1</sup>Referring to the community of clinician-researchers

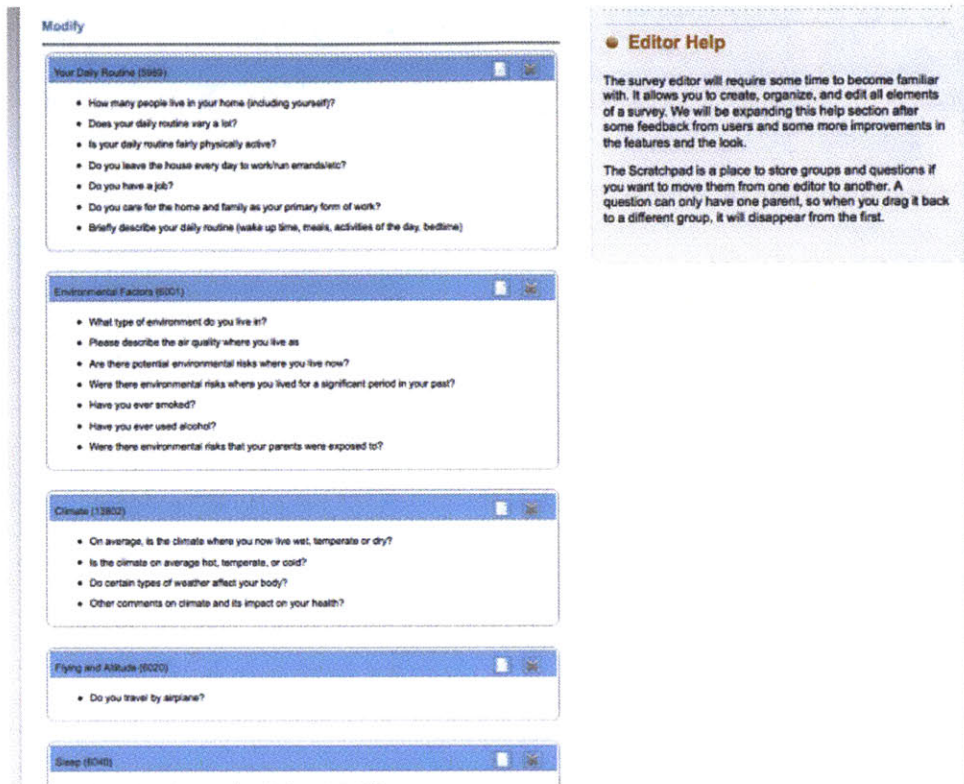


Figure 4-4: LAMsight Survey Editor

it. After a year of advocacy at meetings, and showing the clinicians some of what we were able to do, they started coming up with constructive ideas for how such a site could create value for their work (see 4.4.2).

The framing of the project that we settled on was to argue that our goal was not traditional hypothesis-driven research, but rather supporting hypothesis generation and running formative or exploratory studies. The goal of LAMsight became: “is there enough evidence that an anecdotal phenomenon exists and can be studied?”

Privacy and ethics of patient-directed data sharing was an issue that was often raised during the time I was advocating for LAMsight. I did not keep an accurate record, but of the hundreds of patients who signed up to share information, or that I interacted with personally, only two ever expressed hesitation with sharing their data. Both were still healthy and maintaining professional careers; having a fatal disease was not something they wanted be known outside their immediate family and disease community. Without the



MIT Media Lab brand name and the clear academic mission of the project, the reception with regard to privacy may have been much more negative. However, most of the patients said something similar to: “my life is on the line and if sharing data might help accelerate the pace of discovery, I don’t care about privacy.”

## **4.4 Results**

A small community of patients were active on LAMsight in the early days, but continued engagement was challenging. Most patients lacked the energy to stay engaged on a longer term basis.

### **4.4.1 Patient Questionnaires**

A small group of highly active patients generated a long background questionnaire on women with LAM. The role of hormones in LAM was (and is) not well-understood, despite LAM being one of the few disease that only effects women<sup>2</sup>, principally during childbearing years. Many patient initiated observations had to do with unusual observations about the short term effects of hormones on symptoms of dyspnea (also see 4.4.3).

Patient generated surveys covered topics including:

- Nutrition and exercise,
- Dietary factors,
- Bone injuries and bone pain,
- Pregnancy,
- Alpha-1 antitrypsin deficiency,
- Sexual activity and breathlessness, and
- Alternative therapies.

---

<sup>2</sup>There are rare cases of men with LAM, but they tend to have highly unusual characteristics and many co-morbidities.

Each of these acquired a large number of responses (for the LAM community), for example 25% of LAM patients surveyed (n=25) reported reported breathlessness lasting more than 1 hour after sexual activity, a somewhat surprising result. This may reflect a short-term sensitivity to hormonal fluctuations that is unique to LAM patients. This and other observations led to the study reported in section 4.4.3

## **4.4.2 Researcher Questions**

As the site became more well known within the community, various clinicians approached us with new research questions that they hoped the community could answer.

- One researcher from Finland had been working on Osteoporosis but had a side interest in LAM research. She recognized similarities between the two cell types and wondered if there was evidence of women taking osteoporosis medications having a different disease course than women not taking such medications. Due to the early mortality of LAM, no one database had sufficient numbers of women. I found five patients who were able to engage with the researchers question.
- Another researcher felt that having quick access to a large population of patients would be invaluable during trial design. The ability to communicate and aggregate data quickly would allow researchers without a high volume of clinical exposure to patients to evaluate whether a particular aspect of a trial was too burdensome.
- A third researcher was the father of a daughter with LAM and knew that many LAM patients were taking various off-label medications. Traditional establishments are leery of recording this information. The researcher designed an off label drug survey for the population.

## **4.4.3 A Study of Pulmonary Function**

At a patient meeting in Brighton, UK in 2008, patients reported that a number of them experienced symptoms that varied in concert with their menstrual cycles. At the time, only

limited research had been done on the possible role of hormones in the disease, surprising perhaps in a disease that is not a direct germ-line mutation and only strikes women and only during childbearing years. A researcher specializing in hormones and LAM was present and said that it would be very interesting to know whether this could be validated through objective PFT data, or in subjective variation in Dyspnea (breathing sensation). Further, she wondered if any variation could be correlated to a pre or post ovulation phase to see whether it was associated with estrogen or progesterone, respectively.

I remarked that this would be easy for people to do from home by capturing daily PFT data using home spirometry and recording it in the prototype of LAMsight that I demoed at the meeting. The LAM Treatment Alliance indicated a willingness to fund the spirometers and some of the costs of managing such a study. The "Estrogen Study", as it came to be known, was meant to be a quick, patient-driven process but was delayed due to a variety of factors, including many questions/concerns from the scientific advisory board of the LTA. This friction illustrates the formulaic approach to confirmatory research that permeates biomedicine. Questions that could be easily answered through a formative study of a few patients instead had to be answered ahead of time with limited information. This significantly delayed the start of the study.

Traditional biomedical research is complicated and costly. Much lower-cost research is possible with the emerging landscape of crowdsourced, online research studies, particularly for a large class of questions of vital interest to patients [Swa12b].

## **Study Design**

With support from the LTA, we ran a 5 month study of nearly 30 women with LAM using sub-\$100 home spirometry to record daily FEV-1 and FEV-6 (Forced Expiration Volume) capturing objective variation in lung function. We introduced a new measure into LAM, the Multilevel Dyspnea Profile (MDP) which measured 3 dimensions of patient dyspnea to evaluate subjective changes in breathing perception, and an ovulation kit that for pre-

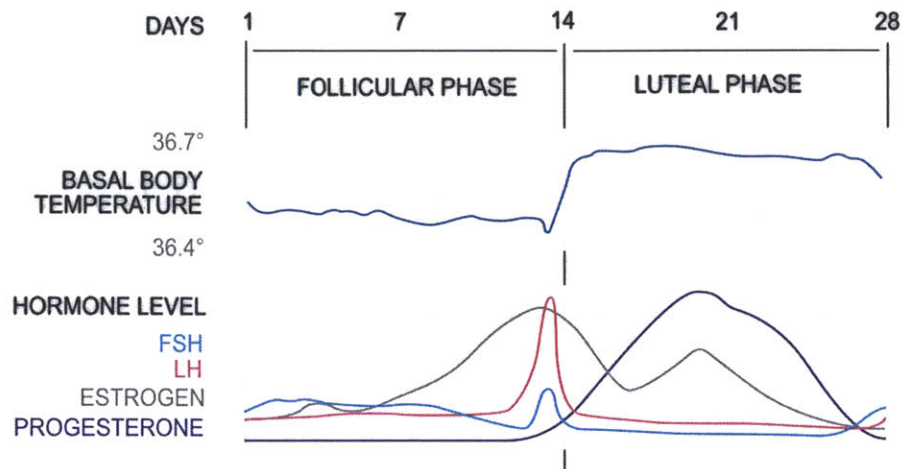


Figure 4-5: Hormone levels during the female menstrual cycle [Iso09]

<b>Standard</b> (n=19)	Women with LAM and menstrual cycles
<b>Complex</b> (n=5)	Women with LAM and unusual conditions (e.g. hysterectomy)
<b>Control</b> (n=5)	Menopausal women or non-LAM women
<b>Withdrawn</b> (n=3)	Women who withdrew early in the study

Table 4.1: Estrogen Study Population

menopausal patients would capture the LH surge to help separate pre-and-post ovulatory hormone activity as shown in Figure 4-5. Basal temperature was a backup measurement for women who reported highly variable cycles. The material cost of the study was less than \$2000, and it was only that expensive because we paid for and shipped a standard set of ovulations kits worldwide.

The goal was to collect a minimum of three months of monthly variation in object and subjective measures, and determine if there was a significant periodicity in the data that exceeded.

## Population

We recruited 30 women with LAM broken into three groups:

Of these groups, seven failed to adhere sufficiently to be included in the final data analysis.

## Study Results

Figure 4-6 provides a visualization of the averaged daily FEV1 data with the self-reported onset of patient menstrual cycles and the LH surge (red and green vertical lines, respectively). Visual inspection shows the co-variation of menstrual onset and the LH surge.

Three women appeared to show the hoped-for variability. In these women, PFT declined during the peak estrogen levels associated with the detected LH surge. The preliminary data was informally presented to several members of the research community. The reaction ranged from skeptical there was any signal, to enthusiastic. The commentary from the latter included:

- This is a unique dataset in LAM, we don't have any data at this scale on short term variation in FEV1;
- The stability of FEV1/6 readings in most patients indicates that home spirometry might be a viable endpoint to track in clinical trials; allowing us to track more frequently and with less burden;
- If the analysis confirms the inverse correlation between Estrogen and PFT data in some women, it suggests there may be a subtype of patient worth of additional analysis;
- If the correlation holds, it may also suggest something about the pathways and mechanisms in LAM; and
- "This data could save lives!"

The formal statistical analysis, publication of preliminary results, and export to LAM experts for secondary analysis were all in progress at the time of this writing.

## 4.5 Lessons Learned

Like many attempts to build social networks outside the community, finding enthusiastic connectors who wanted to engage on the platform was difficult. Moreover, the physical disability of many patients reduced the free time available to engage on a website regularly,

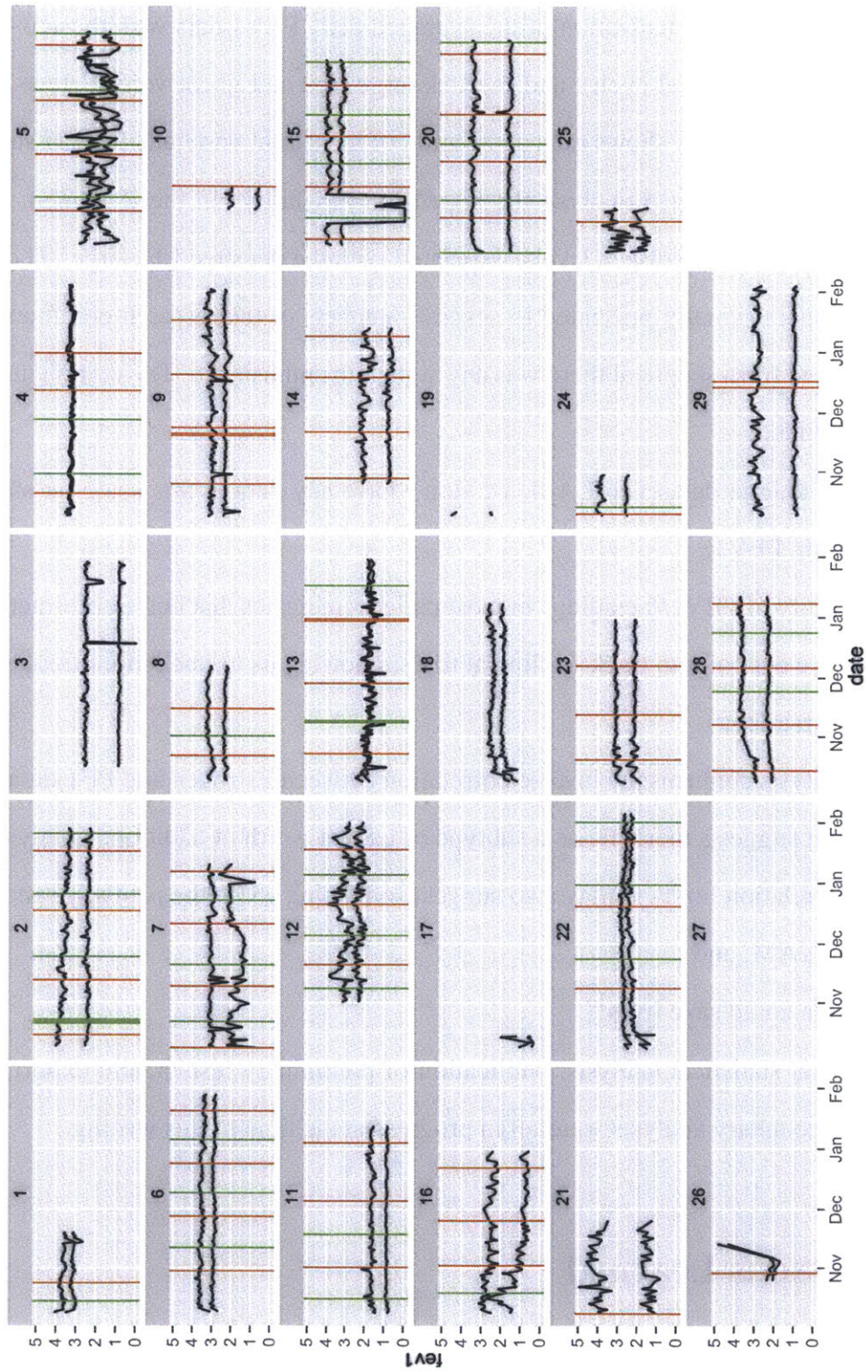


Figure 4-6: FEV1/FEV6 with menstrual onset (red) and LH surge (green)

so our longitudinal data collection ended up being extremely limited in comparison to our one-time background surveys.

The partner organization had committed staff time but it took several years to get an appropriate hire in place given their other obligations. As a consequence, we failed to capitalize on the early excitement and press around the site concept and to keep the content engaging and dynamic. The site itself developed slowly due to variability in partner funding and competing demands on the primary developer. The lack of regular updates and content changes may have contributed to the lack of ongoing engagement.

As a learning exercise, however, the LAMsight deployment was quite successful. Web-based platforms support elicitation and answering of new questions posted by researchers and patients and can facilitate novel forms of research that might not otherwise be funded, generating new ideas for a disease. The experience of LAMsight bears many similarities to a much larger and better funded platform, PatientsLikeMe, described in more detail in Section ??.

There were a number of take away lessons for people looking to implement new platforms, and the best suggestions for future improvement on the current platform came from patients.

1. **Long-term, longitudinal data collection may be unrealistic.** Longitudinal data collection, without the context of a specific study or question, appeared unlikely to reach sufficient volumes in small populations to be statistically meaningful.
2. **Don't try to create a new community if scale is important.** The world of social networking, online data sharing, and even clinical registries has changed dramatically since LAMsight was conceived. Opportunities exist today to integrate with existing communities, using existing authentication mechanisms to reach out for patients. New facilities can be tested as a component (e.g. Facebook app, widget, API-enabled data store, etc) of a larger ecosystem.

3. **Use story-driven outreach.** Patients said they would take time to participate in the pursuit of a specific goal, or to help someone, and probably only if reminded. They all acknowledge that regular, voluntary engagement, was unlikely for most of them.
4. **Identify the 'expert patients'.** A clinician and I decided that a very interested research study would be to ask patients in a visit to predict their test results before the tests occurred and to compare them to the objective data to see if there is a sub-population that can reliably predict their own results. Are there other ways to identify and leverage the expertise of these kinds of patients akin to the Delphi methods [Bro68]?
5. **Subjective endpoints.** Novel at the time, although now becoming mainstream ([FOV<sup>+</sup>11] [WMKD11]), was the use of quality of life and other patient reported outcome measures as primary endpoints of a therapy. Researchers haven't always investigated the phenomenon that concern patients day to day. This is a massive unmet opportunity.



## **Part II**

# **Aggregated Self-Experiments**



# Chapter 5

## Extended Example

This chapter provides an idealized, narrative introduction to the Aggregated Self-Experiments framework using a hypothetical patient case series. The cases are an amalgam drawn from real-world experiences reported on the forum TalkPsoriasis [?] as well as from personal discussions, interviews, and the experiences of the author. It illustrates how the framework functions, by highlighting the decisions a patient makes along with the information and support they need to make them. This idealized example provides a concrete backdrop for the generalized descriptions of the framework in Chapter 6 and Chapter 7.

### 5.1 Case Review

Alice is a 30 year old female diagnosed with psoriasis, an auto-immune disease. The primary symptoms of the disease are patches of inflamed, scaling skin on her torso, scalp, and sometimes arms and legs. Alice also reports feeling severe chronic fatigue and mental “fog,” the severity of which she says varies with her diet and seems to improve when her skin improves.

Psoriasis is a common auto-immune condition found in nearly 2% of the global population [CK03]. It is characterized by skin lesions and often appears after an infection or periods of elevated stress. Though primarily treated by dermatologists, emerging research

indicates that psoriasis is a systemic auto-immune / inflammatory illness associated with an increased risk of premature death. The etiology of the disease remains unknown. Up to 30% of psoriasis patients will develop psoriatic arthritis which manifests as pain, swelling, and stiffness in the joints. Alice has read in online forums that recent studies show significant co-morbidities for more serious cases including a 58% increase in heart disease risk and a 43% increase in stroke risk [Nat12]. This adds up to a 50% increase in the risk of death and 4-year decrease in life expectancy on average [GTL<sup>+</sup>07]. Patients diagnosed under the age of 15 may die as much as 10 years earlier than a demographically equivalent patient without psoriasis.

Alice has a mild case of psoriasis, and feels she can function normally. She appears healthy to others except during flares, when multiple lesions will appear on her face, scalp, and extremities. Flares are typically triggered by stress, alcohol, and sometimes weather. Alice has not opted for systemic immuno-suppressant biologic medications which are the standard of care for more severe disease, although she occasionally applies a steroid cream to control symptoms during flares. She says that performing in her job at times is quite difficult due to the fatigue and mental fog. Thus, while she is concerned about the long-term health risks, her primary concern is reducing the mental symptoms that appear linked with her psoriasis symptoms.

Her blood tests are all normal. She is physically healthy, and runs 4-6 miles once or twice a week to stay in shape. Her primary care doctor said that biologic drugs are not necessary and steroid creams should help with symptom management. However, when she described her fatigue symptoms, he shrugged his shoulders and said that modern medicine doesn't really know how to help her with those kind of secondary symptoms and that there "isn't any data for or against any of the therapies like diet." At the end of her visit, the doctor concluded: "You are very healthy, come back in a year for your next checkup." As she walked out, she said to herself, "how can I be healthy if I report feeling miserable and only partially functional much of the time?"

Her case is not unique. A recent study of the online patient form, TalkPsoriasis, indicated that nearly 75% of patients go to social media sites to get treatment guidance and advice they can't get anywhere else [IR].

## 5.2 Seeking Help Online

Alice regularly searches online for information about her condition. She is a member of TalkPsoriasis.org, an online patient support group sponsored by the National Psoriasis Foundation. Alice has been a member for several years and receives daily digest e-mails, but only occasionally reads them. She finds reading the forums depressing as she fears that the stories of more severe conditions may describe her own future.

Upon starting a new job and relationship, Alice decides to become more engaged in improving her health and trying to better manage her psoriasis. She observes that stress, diet, and alcohol are clear triggers for her condition and she believes that reducing the frequency of these triggers might reduce the number of bad days. She returns to TalkPsoriasis.com and peruses the "Complementary and Alternative" section for commonly reported treatments. In the first hour, she identifies the following list of recommendations:

Dead sea bath salts, Slippery elm bark tea, topical Glycerin and Witch Hazel, Pagano diet, Acupuncture, Alkaline diet, Vitamin B complex, high-dose Vitamin D, Liver Aid, Juicing, Juice fasts, Homeopathy, Olive leaf, Humidifiers, Colonics, Coconut oil, Zinc, Aurveda, Turmeric, Marshall Protocol, Manuka honey, Anti-candida treatment, Koronjo Oil, Broth fasting, Negative ion therapy, Chelation Therapy, Apple Cider Vinegar, "liver flush," Papaya, Meditation, Tea tree shampoo, and White tip Oolong Tea.

Forum users are divided as to whether any of these treatments worked, some reporting

yes and others no. She has trouble finding discussions specifically related to her symptoms of fatigue and mental fog, so asks her question directly in a forum post (transcript in Appendix A.3). Within two days, she receives the following recommendations from 7 distinct users:

```
Avoid triggers (e.g. corn, wheat, dairy), Sulfasalazine,  
Pagano Diet, Furhman Diet, Kale smoothies, Liverite  
liver cleansing product, Milk Thistle and Dandelion,  
Thyroid / Adrenal supplements, Acupuncture and  
Chinese Herbs: ``Dr Shen's Good sleep and worry free,``  
Vitamin B supplementation (3 times), Vitamin C, and  
Selenium
```

She is overwhelmed by the number of options. The Pagano diet [Pag08] is recommended on both lists, so that seems like something reasonable for her to try. During her web searches, she encounters a site called PersonalExperiments.org with a page that documents the Pagano diet along with an experiment for trying the treatment out. There is also a record of how many users experienced positive results (Figure 5-2). She registers to use the site.

## 5.3 Personal Experiments

PersonalExperiments.org is a web prototype of the Aggregated Self-Experiments framework [ECL11].<sup>1</sup> Personal Experiments helps users document symptoms over time, using devices or SMS delivery, and to evaluate the impact of treatments through simple experiments run in the context of their everyday lives.

---

<sup>1</sup>Most, but not all, of the features described here are implemented in the prototype. Some described features are idealizations of design goals for future implementation. Chapter 8 describes the specific implementation in more detail.

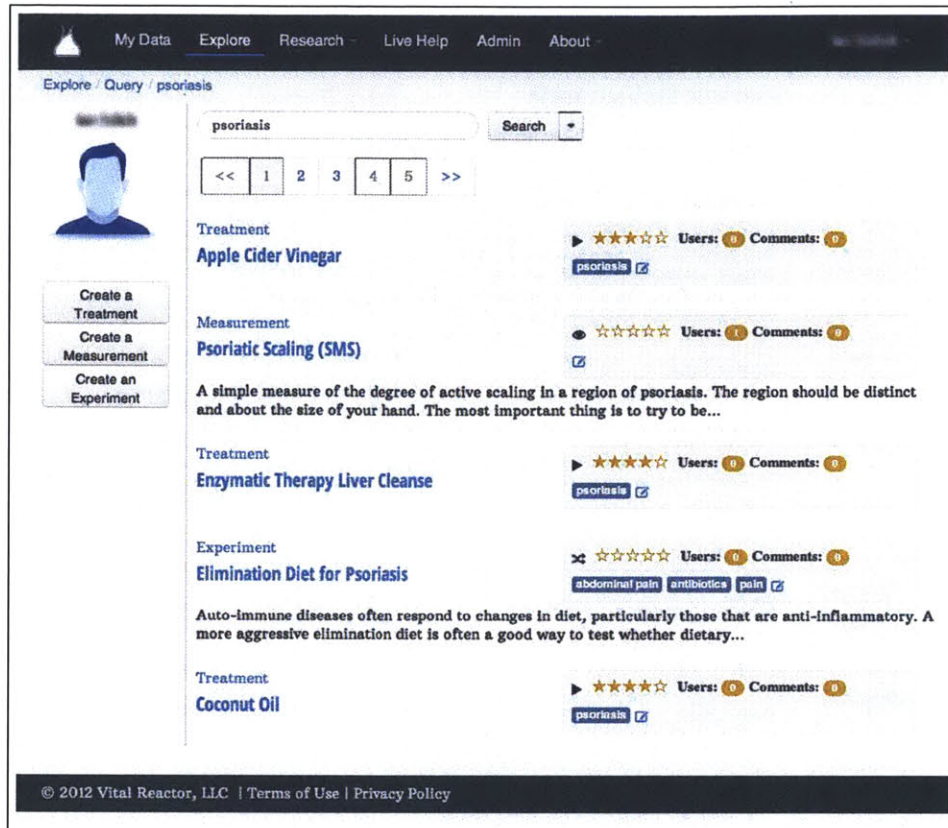



Figure 5-1: Searching for Psoriasis-related Information

The site provides data collection, journaling and data review pages, and a catalog of *Treatments*, *Measurements*, and *Experiments* (Figure 5-2). *Treatments* describe a particular health intervention which may be a habit, a supplement, or a drug, in human comprehensible terms. It also captures information about how the treatment behaves during transition periods between a treated and untreated state (i.e. onset and carryover effects).

*Measurements* describe a mechanism by which the system will measure a *Variable*. A *Variable* is a user-defined text label referring to an underlying, measurable quantity such as total sleep, nightly wakings, or systolic blood pressure. There may be multiple *Measurements* defined for a given variable defining how the value of the measure is acquired.

The default mode of measurement consists of scheduled short prompts that are delivered and collected via SMS, but it also provides support for entering data directly on the web site. Other measurement types supported include native mobile apps, 3rd party web

Explore / View / Measurements / Psoriatic Scaling



See Profile

- Create a Treatment
- Create a Measurement
- Create an Experiment

## Psoriatic Scaling (SMS)

Untrack Edit

Details Discussion

### Description

A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. The most important thing is to try to be consistent in what the psoriasis looks like from measure to measure.

### Service Type

SMS or Email Messaging

### Prompt Text

Please rate the amount of scaling from 0-4 where 0 is normal skin, and 4 is thick heavy plaques.

★★★★★ Users: 1 Comments: 0


External References

New Reference

### Related

- ⌕ Glycerin and Witch Hazel for Psoriasis
- ⌕ Elimination Diet for Psoriasis

Explore / View / Treatments / Glycerin and Witch Hazel



See Profile

- Create a Treatment
- Create a Measurement
- Create an Experiment

## Glycerin and Witch Hazel

New Experiment Edit Clone

Details Discussion

### Description

A topical treatment some patients find help with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied.

Impact on scaling in psoriasis can be seen in responding patients within two weeks, often sooner for other symptoms such as redness and scaling.

### Side Effects

### Behavior

days to take effect: 14

days to wash out of your system: 7

★★★★★ Users: 1 Comments: 0

psoriasis

External References


- Patient Hypothesis
- Patient Recipe and Places to Buy
- 2003 Research on Glycerin for Skin

New Reference

### Related

- ⌕ Glycerin and Witch Hazel for Psoriasis

Explore / View / Experiments / Glycerin and Witch Hazel



See Profile

- Create a Treatment
- Create a Measurement
- Create an Experiment

## Glycerin and Witch Hazel for Psoriasis

Stop Trial Edit Clone

Details Discussion

### Hypothesis

Topical or oral Glycerin and Witch Hazel help re-establish normal barriers between layers of developing skin cells that are disrupted by the inflammatory processes in psoriasis patients.

### Treatment

Glycerin and Witch Hazel

### Outcome

Psoriatic Scaling

Success will **decrease** this measure

### Other Measures

Psoriatic Itchiness

★★★★★ Users: 1 Comments: 0

abdominal pain antibiotics pain

External References

New Reference

### Related

- ⌕ Elimination Diet for Psoriasis

Figure 5-2: Measurement, Treatment, and Experiment Views



services, consumer devices and built-in web surveys. For example, nightly sleep duration can be elicited via an SMS message, which relies on the user to track their sleep manually. Sleep can also be measured directly using an activity tracker such as Fitbit, Jawbone Up or Zeo and downloaded to the site via each device's service API.

An *Experiment* links a treatment to one or more variables. The experiment provides a human description of the experimental hypothesis (what will happen and why) and any additional instructions for how to run the experiment. The experiment has a primary *Outcome* variable used to assess the treatment's effectiveness. The default experiment design (treatment-withdrawal) asks the user to collect data during three time periods: a pre-treatment baseline phase, a treatment phase, and a post-treatment (or withdrawal) baseline phase. The analysis compares the pre- and post-baseline data to treatment data to determine if the change in the outcome under treatment was significant. The data review page provides a trial chart to assess the outcomes, which is described in more detail in Section 8.4.2.

Other supported features include individual data point annotations and an SMS or web-based journaling mechanism for observations that may be relevant to interpretation of the data or peer review of a trial, but are not needed for analytical purposes.

A user creates a *Trial* of an experiment (Figure 5-3) by providing a start date, selecting measurements for the experiment's variables, and scheduling individual *Trackers* for each of those measures.

A user can choose to create one or more standalone trackers for observational, or self-tracking [Wol10] purposes. Each tracker manages a series of values defined by a specific measurement. Trackers vary based on what kind of data source is tracked. For momentary assessment variables, the user can configure the schedule to regularly or randomly prompt for the current value of a variable.

Configured trackers and trials populate a user dashboard (Figure 5-4), a single place where all day-to-day data actions may be recorded. Common actions include recording data, recording journal entries, changing and annotating already-recorded data, and manip-

Advanced View

## Start a trial of: Glycerin and Witch Hazel for Psoriasis

**Trial Instructions**

Apply treatment during the treatment periods only, use regular care in the baseline periods. No other special instructions needed.

**Treatment Description**

A topical treatment that some patients find help with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied.

Impact on scaling, redness and inflamed area in psoriasis can be seen in responding patients within two-three weeks, often sooner for other symptoms such as redness and scaling. Itching often subsides within days.

**Potential Side Effects Glycerin and Witch Hazel**

Witch hazel with alcohol can sting on application. Application of this mixture to open sores (e.g. after scratching) can hurt a little. No other harms or side effects are known.

**Configure your trial**

When do you want to start?

13
March
2013

Do you want treatment reminders?

Yes  No

**Select your measurements**

**Psoriatic Scaling**

via sms

**Psoriatic Itchiness**

via sms

**Psoriatic Redness**

via sms

**Configure trackers**

**Reminders for 'Glycerin and Witch Hazel for Psoriasis'** ⊙ || ⊙

Please don't forget to apply your glycerin/witch hazel mix

Reminders via SMS Daily at 10:00 am

**Psoriatic Scaling (SMS)** ⊙ || ⊙

A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. The most important thing is to try to be consistent in what the psoriasis looks like from measure to measure.

Create
Cancel

Figure 5-3: Configuring a Trial

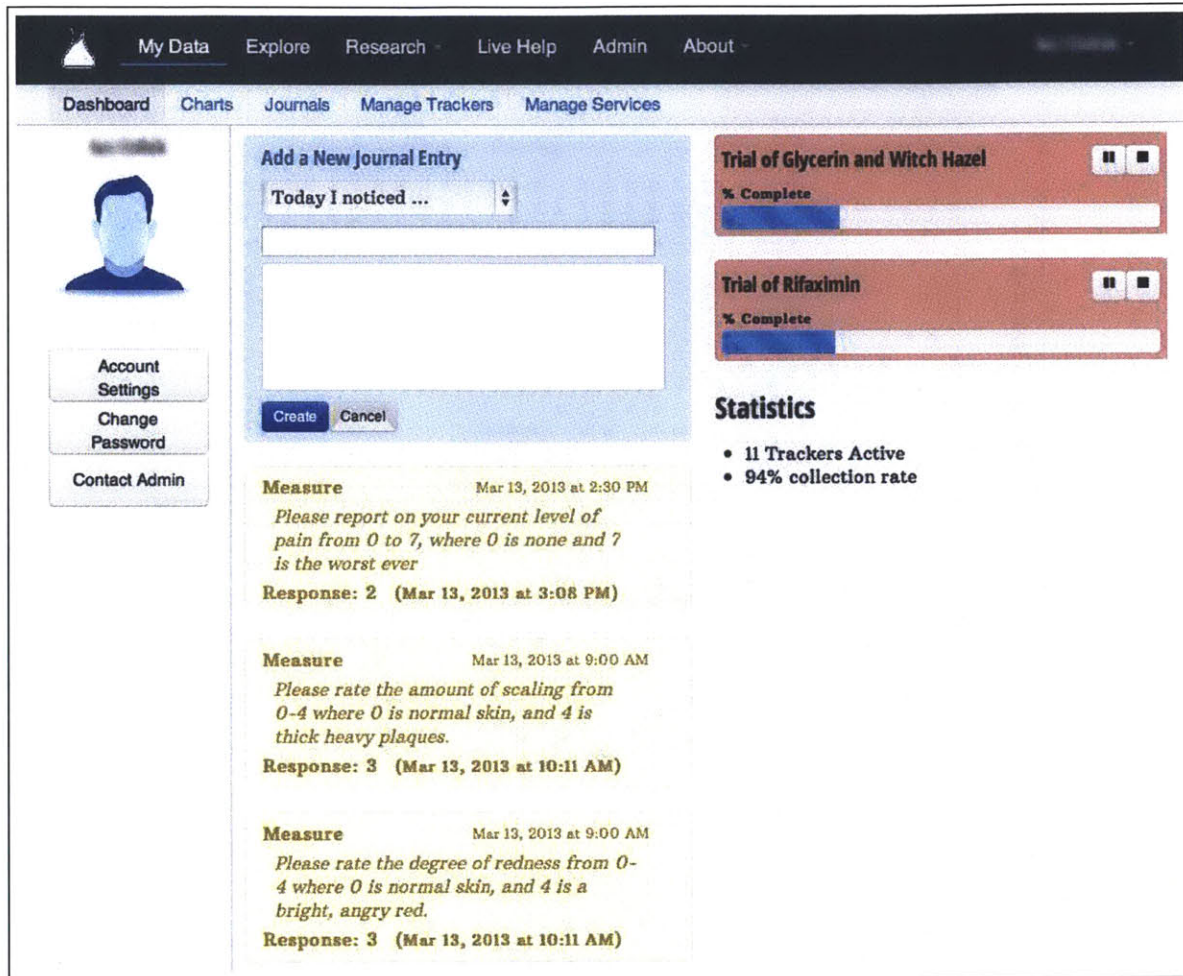


Figure 5-4: User Dashboard

ulating active and completed trials.

To review in-progress or post-trial data, the user clicks on the charts tab where they can see two kinds of data summaries: trial timelines and tracker timelines (Figure 5-5). A calendar axis, journal glyphs, and medication bars provide contextual information.

The tracker chart plots all the observed data points along with the mean and upper and lower “control limits” that characterize the range of normal readings of the measure.<sup>2</sup> Points that fall outside the control limits are circled on the chart. The user’s interpretation is that “something interesting happened” at that point.

For active trials, the primary outcome variable is charted against a shaded background

<sup>2</sup>Section 7.2.6 describes control charts and limits in detail.

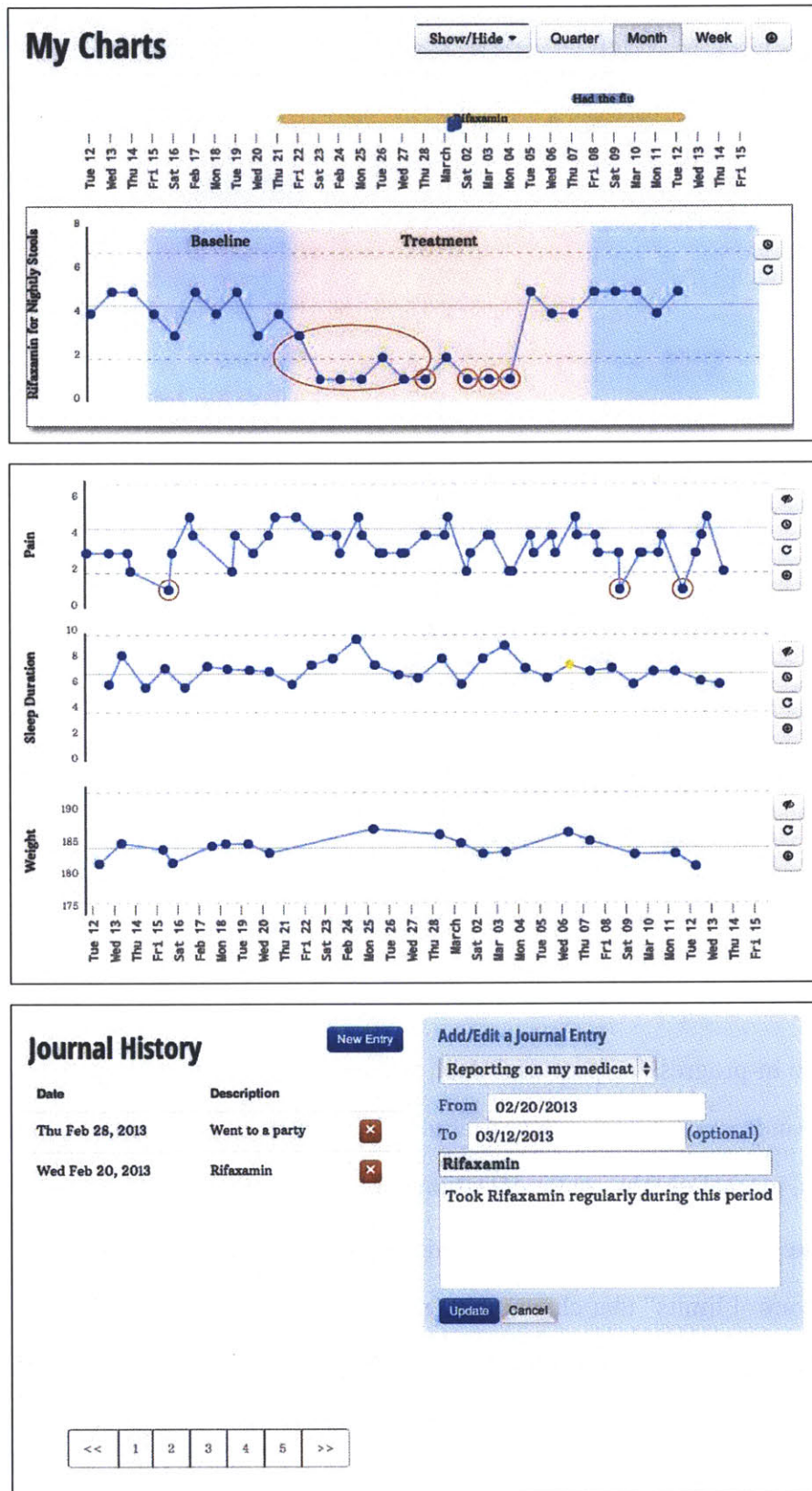


Figure 5-5: Trackers, Trial Control Charts, and Journaling

for each of the different trial phases (Figure 5-5). The control limits are computed only from the baseline periods and the significant events are computed for the treatment periods. If treatment periods produce significant events, then the experiment is determined to be a success. Control chart computations are driven by a statistical model of the trial described in Chapter 7.

If a user gets a negative result, but something unusual happened during the trial, they can disagree with the result of the trial and identify the major reason why they think it didn't work, opting to extend or retry the study. Section 8.4 discusses other actions that can be taken when special causes interfere with proper interpretation of a trial.

## **5.4 Alice's Trials**

Alice notices on the search page a "Recommended for me" button. A dialog box asks for the symptoms she is interested in improving and background such as her diagnosis of psoriasis. She is shown a list of ranked experiments to try. The order of the experiments reflects how likely each is to work for her. The Pagano diet is not ranked highly as a means to reduce symptoms of psoriasis. A simpler treatment, glycerin and witch hazel (Figure 5.1), is at the top of the list as working for over 90% of the people who try it. The experiment only takes two weeks, whereas the Pagano diet takes several months.

### **5.4.1 Glycerin and Witch Hazel for Psoriasis**

There are two experiments defined for the glycerin and witch hazel treatment. The one ranked highest in her results uses itching as the outcome variable. The treatment works for most people in reducing itching. Alice selects a lower ranked one that uses psoriatic scaling as an outcome variable. This particular trial is documented by links to clinical trials on the potential role of glycerin in stabilizing the formation of skin cells in psoriasis patients [ZRB03]. The protocol is summarized in Table 5.1.

<b>Title</b>	<b>Glycerin and Witch Hazel for Psoriasis</b>
<b>Hypothesis</b>	Topical or oral Glycerin and Witch Hazel help re-establish normal barriers between layers of developing skin cells that are disrupted by the inflammatory processes in psoriasis patients.
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Redness, Psoriatic Itching
<b>Design</b>	1 week baseline, 2 weeks treatment, 2 weeks baseline
<b>Treatment Description</b>	Glycerin and witch hazel is a topical treatment that patients find helps with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied. Impact on scaling in psoriasis can be seen in responding patients within two weeks, often sooner for other symptoms such as redness and scaling.
<b>Side Effects</b>	None reported.
<b>Protocol</b>	Choose one area of your body to treat for this study, treating the rest of your body as you normally do until the end of the trial. This will help you compare and contrast your normal habits with the treatment. Start by measuring normal psoriasis activity for a week, apply the treatment for two weeks, then remove the treatment again to see if symptoms return without treatment for another two weeks. <b>WARNING:</b> For some people symptoms do not return, in which case the trial may report a negative outcome in error when in fact it was a success. The trial may be updated in the future to account for this case.

Table 5.1: Experiment Detail for Glycerin and Witch Hazel

Three common, co-varying symptoms of interest for this treatment are Scaling, Redness, and Itching. The measurement methods available for these variables are SMS- or web-based prompts:<sup>3</sup>:

- **Scaling** - “Rate the magnitude of scaling from 0-4, where 0 is no flaking and 4 is thick plaques”
- **Redness** - “Rate the magnitude of redness from 0-4, where 0 is normal skin and 4 is angry red”
- **Itchiness** - “Rate the magnitude of itchiness from 0-4, where 0 is no itching and 4 is can’t stop”

Alice selected treatment reminders to be sent every day during the trial, then clicked Start. The trial added a progress summary widget to her dashboard that tells her about her

<sup>3</sup>Derived from a psoriasis assessment instrument developed by the National Psoriasis Foundation [Fel05] [GCB+03]

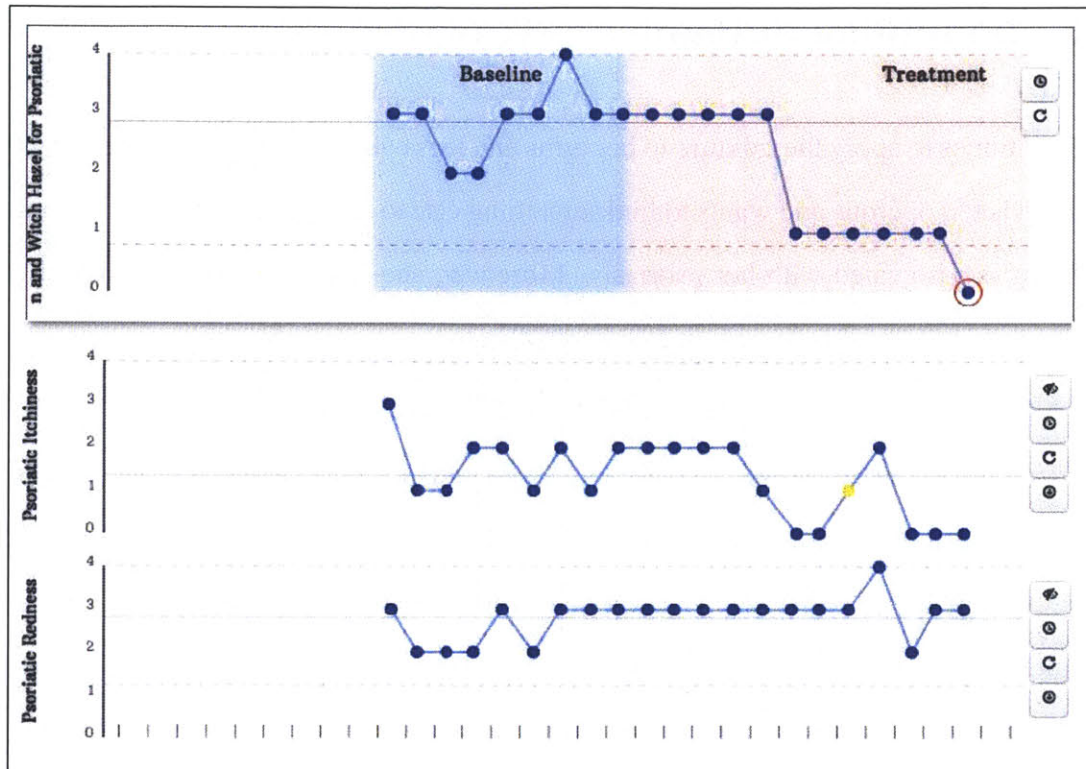


Figure 5-6: Glycerin and Witch Hazel, Intermediate Trial Results

% progress into the trial. She begins receiving SMS prompts the next day in the morning and evenings, occasionally logging in to her dashboard to enter the previous day's data when she forgets to respond. Her adherence to data collection is nearly perfect.

## Result

Within 2 days of starting treatment, Alice's normal level of low grade itching goes from a baseline average of 2.5 to 0, her scaling declined from 3.5 to 1 on day 5 of treatment but the redness was largely unchanged during the entire treatment period (see Figure 6-1).

After treatment, the scaling starts to slowly return to baseline along with the itchiness. She begins applying the treatments again after that first week and logs in to stop the trial. The system asked her why she is stopping and she picks a response indicating that the treatment clearly worked and she is continuing it rather than waiting for the end of the trial.

## 5.4.2 Turmeric for Psoriasis

Alice continues to apply the mixture to her arms and legs every day, but feels that doing this long term is too onerous and wants to find something else to stop the systemic inflammation she thinks is associated with her psoriasis. Moreover, she is still fighting the fatigue and mental fog symptoms that first motivated her to turn to the internet. She believes that topical treatments are unlikely to help with systemic symptoms.

Dietary interventions appear to work in 15% of patients, but she doesn't feel capable of making the necessary changes due to her travel schedule. Encouraged by her earlier experience, Alice searches on Personal Experiments for a simple systemic treatment. Turmeric, a supplement, is reasonably inexpensive, easy to take and improves symptoms for about 20% of patients when taken over a longer term period.

### A Special Cause

A few days into the treatment phase for turmeric, Alice engages in several days of drinking which leads to a flare-up. Her mental fatigue increases so much during this period that she forgets to take her Turmeric and ignores the SMS prompts on her phone. After ignoring prompts for a few days, the site sends an e-mail asking whether she is still interested in continuing the trial or changing it. She clicks the link and selects "Restart Phase" to restart her abandoned treatment phase. When asked about the reason for restarting, she clicks "Got Sick".

The site then asks for data for four days, indicating no treatment should be taken. After, it asks her to restart taking Turmeric<sup>4</sup>.

---

<sup>4</sup>This calculation is based on the desire to re-establish whether current measures are consistent with previous baseline data, or whether more baseline data is needed. Variance of the measure, anticipated washout, contributed to the calculation discussed in Section 7.3.



<b>Title</b>	<b>Turmeric to reduce Scaling</b>
<b>Hypothesis</b>	Curcumin, the active ingredient in turmeric, is effective in slowing down the energy supply to the rapidly dividing cells that result in red, scaly, skin symptoms. Taken as a dietary supplement, Turmeric or Curcumin can help to improve the appearance of red, scaly, inflamed skin.
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Redness, Psoriatic Itching
<b>Design</b>	2 weeks baseline, 3 weeks treatment, 2 weeks baseline
<b>Treatment Description</b>	<p>Turmeric (<i>Curcuma longa</i>), the major ingredient of curry powder and prepared mustard, has a long history in both Chinese and Ayurvedic (Indian) medicine as an anti-inflammatory agent. An increasing number of recent in-vitro, animal, and human studies of turmeric have shown positive outcomes by slowing down the rapid skin cell turnover that occurs in patients with psoriasis.</p> <p>The volatile oil fraction of turmeric has demonstrated potent anti-inflammatory activity in a variety of experimental animal models, while curcumin, the yellow pigment of turmeric, is even more potent in acute inflammation.</p> <ol style="list-style-type: none"> <li>1. When used orally, curcumin inhibits leukotriene formation, inhibits platelet aggregation, and stabilizes neutrophilic lysosomal membranes, thus inhibiting inflammation at the cellular level.</li> <li>2. Curcumin is reported to possess greater anti-inflammatory activity than ibuprofen.</li> <li>3. At low levels, curcumin is a prostaglandin inhibitor, while at higher levels it stimulates the adrenal glands to secrete cortisone.</li> <li>4. Formulation difficulties due to the yellow color of curcumin have made topical use slow in coming. However, recent developments in technology may change that. The standard oral dose of curcumin is 250-400 mg, three times a day.</li> </ol>
<b>Side Effects</b>	<p>Turmeric is known to act as a blood thinner or anti-coagulant. Hence this should not be taken with other blood thinning medicines. Some examples of blood thinning medicine are aspirin, clopodogrel, enoxaparin, heparin etc.</p> <p>Turmeric should also be avoided by those who have congestive heart disease, gallstones, liver disease, jaundice or acute bilious colic.</p> <p>In rare cases turmeric could cause diarrhea, sweating and nausea.</p> <p>There is not enough research to establish the risks of very high intake of turmeric and hence it is best to stick to the recommended dose.</p>
<b>Protocol</b>	The easiest option to get the required dosage of turmeric is through dietary supplements. The suggested dosage of turmeric for psoriasis is extracts containing at least 500mg of curcumin, three times a day for 1500mg/day.

Table 5.2: Experiment Detail for Turmeric

## **Result**

After three weeks of taking Turmeric, she feels better, but after stopping turmeric doesn't notice any change. At the end of the final baseline phase, the system sends her an e-mail report of her trial outcome indicating that her response to turmeric was indeterminate. The system indicates that the two baseline phases were too different (the variance of the second phase being much higher than the first) and suspects that something else was influencing her outcome. It says that continuing the study might yield a more definitive result. Since the effect she thought she felt wasn't strong enough to redo the experiment, she goes back to work and forgets about Personal Experiments for awhile.

Unbeknownst to both her and Personal Experiments, she had been careful to avoid all alcohol after the martini incident given the recent harsh experience, but started having occasional glasses of wine on date nights again during the second baseline period. These wine nights would cause a few days of psoriasis activity, increasing the variation over what she experienced during her first baseline period where she hadn't been going on many dates.

Both the interruption of the treatment period and the increased wine drinking in the 3rd phase of the trial are examples of confounding factors (Section 2.5) that must be accommodated in trials. Confounding factors are influences that change the outcome measure independent of the treatment. A weak decrease in psoriatic scaling between the 1st baseline and 2nd treatment phase could have been caused by either turmeric, the extended absence of alcohol, or a third unknown factor. The introduction of alcohol confounded the baseline measures, indicating that something other than the treatment influenced her measurements and obscured the modest treatment effect.

### **5.4.3 Restriction Diet for Fatigue**

Several months later, Alice's symptoms are not bothering her much due to the use of glycerin to control skin symptoms. She finds out she will not travel for awhile and decides to try

working on her diet to address her mental health issues. While she is good about avoiding two known triggers, alcohol and sweets, she otherwise eats a standard American diet.

Alice returns to Personal Experiments and the Pagano diet page. While reading user comments on the Pagano diet, another user suggests that she start with a shorter, more restrictive dietary experiment. While only some people benefited from dietary manipulation, the site indicates that people who respond to a highly restricted diet or fast also respond to more sustainable diets like Pagano. The major benefit of this step is that restriction diets or fasts have shorter onset times than the Pagano diet.

The experiment is described in Table 5.3. Alice tracks her psoriasis symptoms throughout the trial to see if there is any change in those symptoms, despite the trial being designed primarily to test changes in fatigue.

## **Result**

As described in the protocol, Alice experiences severe cravings during the trial. She sticks with it and finds that at 2 weeks her energy level increases dramatically and it is noticeable to everyone in her life. She almost hates to stop at 3 weeks and re-introduce her usual foods. The return of fatigue after starting to eat her normal foods was dramatic. Within 3 days she decides to retain at least the gluten- and dairy-free aspects of the diet and focus on adding back more starches, sugary foods, and the occasional dessert. Her fatigue returns strongly enough during the final baseline phase for the trial result to be clearly positive.

Aside from the improvements in energy, she sees evidence of her psoriasis clearing as healthy skin starts appearing in the middle of inflamed patches, promising evidence that dietary changes help both mental and physical symptoms.

### **5.4.4 Pagano Diet for Psoriasis**

After her dramatic experience during the restriction diet experiment, she goes online on TalkPsoriasis to ask others about their experiences. The general opinion is that restricted

<b>Title</b>	<b>Restriction Diet for Fatigue</b>
<b>Hypothesis</b>	Patients with auto-immune disease often experience fatigue as a by-product of inflammatory processes triggered by specific foods. Removing all trigger foods and gut-unfriendly substances like refined sugar significantly reduces auto-immune response.
<b>Outcome</b>	Fatigue
<b>Covariates</b>	Total Sleep
<b>Design</b>	2 weeks baseline, 3 weeks treatment, 2 weeks baseline
<b>Treatment Description</b>	<p>This diet minimizes sugar and other potential triggers of immune response such as gluten, dairy, and nightshades. It has been used successfully by some patients to reduce the severity of psoriasis and causes partial or full clearing in some.</p> <p>The restriction diet guidelines include:</p> <ul style="list-style-type: none"> <li>Gluten free (no wheat-based products, try alternative grains)</li> <li>Low-sugar (no processed sugar, limit fruits to 1-2 servings a day)</li> <li>Remove triggers: chocolate, strawberries, tomatoes, nightshades, etc.</li> <li>Limit consumption of red and processed meats (1-2 servings / week)</li> <li>Omega-3 supplements or more fatty fish (salmon, mackerel, sardines)</li> </ul> <p>Examples of food that can be eaten:</p> <ul style="list-style-type: none"> <li>Poultry, fish, and eggs</li> <li>Berries in moderate quantities are acceptable</li> <li>Any vegetables not on the trigger list</li> <li>Pastas and cereals made from potato or rice flours</li> <li>Use unsweetened Almond Milk instead of Milk for cereal, shakes</li> <li>Low-lactose milk-based products such as hard cheeses are fine</li> <li>Healthy fats such as olive oil or flax seed oils for flavor</li> <li>All nuts</li> <li>Large salads with beans, chopped veggies, and oil-based dressings are great; can add canned salmon or chicken for protein</li> </ul>
<b>Side Effects</b>	Some people will experience intense sugar cravings for the first 2 weeks on the diet. Most people think this is a good sign, as cravings are generally a sign that you will respond to the diet. The more disciplined you are about sugar removal, the shorter the cravings will last.
<b>Protocol</b>	Follow the diet during the treatment phase. Returning to your standard diet can be hard towards the end of the trial, but it's important to stick with it for two weeks to confirm that it was the diet that was causing your improvement.

Table 5.3: Experiment Detail for the Restriction Diet

<b>Title</b>	<b>Pagano Diet for Psoriasis</b>
<b>Hypothesis</b>	Pagano believes that psoriasis stems from damage to the intestinal tract from eating the wrong foods, leading to irritation of the intestinal lining. Over time, the irritation causes the intestinal lining to become thin and unable to effectively screen out toxins and large food particles. As a result, toxins build up in the bloodstream, taxing the ability of the liver and kidneys to process them. As a result, they exit through the skin. In some people, according to Pagano, this combination of events causes the manifestation of psoriasis.
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Redness, Psoriatic Itching, Fatigue
<b>Design</b>	2 weeks baseline, 4 months treatment, 1 month baseline
<b>Treatment Description</b>	<p>Drink 6-8 glasses of water daily</p> <p>Avoid all processed foods, including pickled and smoked foods or those containing coconut or palm oil.</p> <p>Avoid shellfish and nightshade</p> <p>One to two servings daily of stewed fruit (for limitation)</p> <p>Psyllium husk for fiber</p> <p>One tsp slippery elm bark powder per 1 cup boiling water</p> <p>Supplement of burdock root, sarsaparilla, yellow dock and beet juice</p>
<b>Protocol</b>	Simply follow the daily routine of supplementation and avoiding foods on the list and anything you have identified on your own that are triggers. Reducing response during the treatment should accelerate your response.

Table 5.4: Experiment Detail for the Pagano Diet

diets help with symptom control by avoiding aggravation of the underlying auto-immune condition. Some people found that so-called “healing diets” increased their robustness to violations of their dietary restrictions. Alice decides to try the Pagano diet experiment for psoriasis (Table 5.4) while maintaining some of the restrictions of her prior experiment (gluten and dairy).

## Result

The results of this experiment are much less dramatic, and Alice has trouble adhering to the treatment regularly. The data during her treatment period clearly shows short term bursts of fatigue and flares of psoriasis after ingesting alcohol or other trigger foods. This, however,

is helpful to her in making ongoing changes to her diet. Overall, there seems to be some effect to the Pagano diet as her fatigue measurements slowly improve. Her psoriasis never clears, but it does decrease in severity, restricted to parts of her body not exposed to the summer sun. The trial result is a success, but with a low effect size.

## **5.5 Six Months Later**

After the Pagano trial, Alice continues to adjust her diet to avoid processed foods, limit alcohol, and other triggers she has identified over time. Now that she is feeling much better, the short term reaction to drinking and other triggers is much more obvious. She continues to use PersonalExperiments to measure fatigue and note when she eats something that might have an effect so she can remember. Noticing the correlations between peaks in her fatigue and what was journaled a day or two before has been helpful to her.

One day Alice receives an e-mail from the site asking her if she is still using glycerin and witch hazel or turmeric and whether she still considers them effective. The e-mail says this will help other users decide whether to use one or both treatments. She says yes to both on glycerin and witch hazel and no to using turmeric but yes to believing that it is effective. She feels that given more time turmeric might have been modestly helpful to her.

One year after first digging into TalkPsoriasis Alice is managing her condition, feels good most days, and is actively learning how to live a healthier life within the reality of her auto-immune condition. She feels a sense of control and hopes that these changes, by reducing daily inflammation, help reduce the long-term risk of life-shortening co-morbidities that are associated with psoriasis.

# Chapter 6

## Framework

This chapter builds on Alice’s narrative to introduce a generalized framework of Aggregated Personal Experiments. It details the processes, decisions, and supporting information needed to support a scientific process of inquiry at both the individual and population level. The framework provides a starting point for defining the formal algorithmic building blocks of Chapter 7 that supports experiments like Alice’s and the design of the dissertation prototypes detailed in Chapter 8.

### 6.1 Hypothesis Space

Informally, a user is trying to find a treatment that will help improve one or more symptoms. This notion can be formalized by considering each trial as a hypothesis embedded in a 3-dimensional space of  $U_n: \{ \text{Interventions}^1 \times \text{Outcomes} \times \text{Designs} \}$ . A design includes all the constraints (number of phases, onset/washout periods, etc.) that dictate a specific means of evaluating whether the intervention improves the mean value of a measured variable by a clinically significant<sup>2</sup> amount. The same intervention-variable pairing may be tested by

---

<sup>1</sup>In this chapter I switch to using the more general term “intervention” instead of a more user-accessible notion of “treatment.”

<sup>2</sup>Except for some specific discussions, I use the term “significant” to mean an effect size that would matter to a patient, rather than the statistical notion of significance represented by p-values.

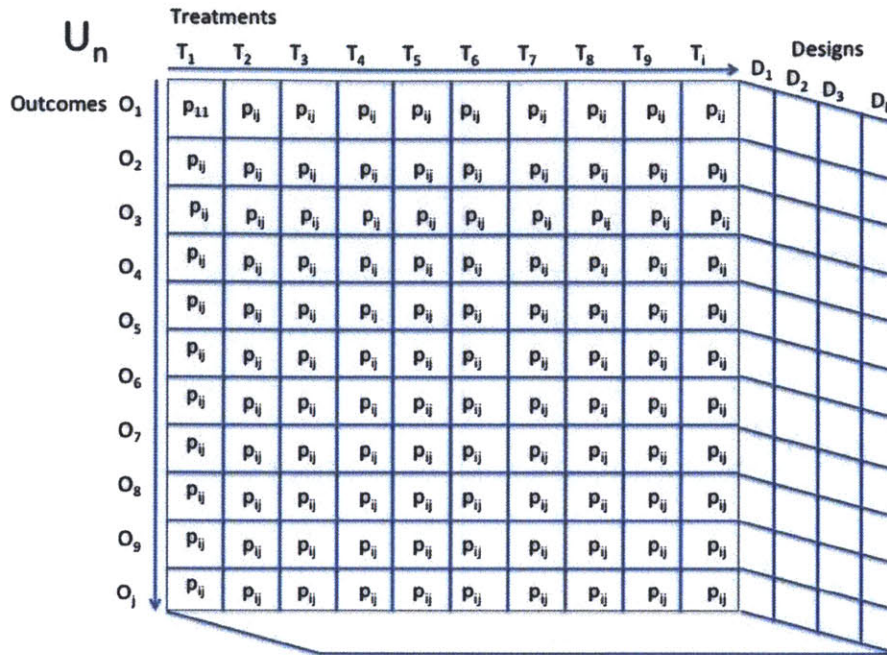


Figure 6-1: Hypothesis space for a single user

different designs. The quality of the design will influence the reliability of the experiment.

How large is this space in practice? In section 5.2, Alice found that an auto-immune condition like psoriasis has dozens of patient-identified interventions in addition to conventional pharmaceutical options. There are a handful of common outcomes of interest, such as fatigue, scaling, redness, and itching in Alice's case. In practice, there are a low-single-digit number of practical different designs. However, there could be a proliferation of concrete designs proposed by users due to disagreements about the pharmacodynamics of the interventions and confounding factors. A user with psoriasis might face a space of hypotheses with hundreds of points. Alice's experiments each took weeks to months, indicating that searching this space naively could take a lifetime of effort. Even searching for treatments over a subspace delineated by a single design style  $T$  and outcome measure could take years.

Under the assumption that successful interventions are likely to exist, the goal of our framework is to minimize the time spent searching for the treatment while maximizing confidence in our beliefs at each step. An action taken in this space represents an exper-



iment that results in an updating of our belief given the result of a Bernoulli trial. The formal representation of belief in our hypothesis space can be captured by a beta distribution at each point in the space. The beta distribution represents our current confidence in the probability of a successful experiment for a given user:  $Beta(\alpha_{uijk}, \beta_{uijk})$ . Without any a-priori knowledge, the beta distribution can be set to a Jeffrey’s uninformed prior  $Beta(\frac{1}{2}, \frac{1}{2})$  to represent an unbiased belief in the probability being between 0 and 1. Without any user-specific prediction or results, the user prior is the same as the population prior:  $Beta_{uijk} = Beta_{ijk}$ .

This formulation of a hypothesis space concretizes the three subproblems that will be the focus of the mathematical machinery described by this dissertation:

- Maximize the probability of selecting the best hypothesis for the next trial;
- Maximize the probability that a trial has a definitive result; and
- Minimize the time required to perform the trial.

The user workflow complements these three goals by facilitating the creation and review of interventions, measures, and experimental designs. The workflow must be carefully designed to constrain the growth of the hypothesis space so that mathematical optimization remains tractable while directing the user behavior to collect data suitable to learn aggregate models. There are additional considerations that the framework must address including user communication, user engagement, treatment safety, and ethics.

## 6.2 Individual Process Model

When a user tries something out in their everyday life, there is often an implicit hypothesis. However, the hypotheses behind “trying things out” is rarely well-specified nor do people regularly employ experimental controls. With the Aggregated Self-Experiments framework, they have access to a pre-defined catalog of treatments and related experiments. Choosing an experiment makes an implicit hypothesis explicit, providing a common struc-

1. **Question** “How can I reduce psoriatic scaling?”
2. **Hypothesis/Prediction** “Glycerin and witch hazel will reduce scaling”
3. **Testing** “Record scaling across periods of treatment and non-treatment”
4. **Analysis** “Was the trial a success? Should I continue/abandon the intervention?”
5. **Iteration** “If my question wasn’t fully addressed, I need a new hypothesis”

Figure 6-2: Alice’s Scientific Method

ture by which two people can meaningfully compare their experiences.

The process that Alice went through in the prior chapter closely parallels the steps of the scientific method (Figure 6-2). The site was able to focus her efforts to facilitate an increased rigor of process that improved her confidence in the outcomes.

A large number of concrete decisions have to be made to test a hypothesis:

1. **Intervention Selection** What intervention do I want to test?
  1. What is the onset and carryover behavior of the intervention?
  2. Does the intervention have practice effects (i.e. needs a longer baseline period)?
3. **Outcome Selection** What primary measure do I want to improve?
4. **Experiment Design** How do I test the hypothesis?
  1. How many phases do I need to account for confounding?
  2. Can/should I blind the intervention?
  3. Is additional value gained by randomization?
  4. What is the sample size in each phase?
5. **Analysis** How do I evaluate the accumulated evidence?
6. **Decision** Do I continue to use the treatment?

- Well-designed experiments of each intervention against a relevant variable.
- A model for computing  $P(\text{Success}|\text{Exp})$  or  $E[\text{Beta}_{ijk}]$ .
- A minimal estimate of time required for the experiment,  $\text{Time}_{exp}$ .
- The probability that success is sustained long term  $P(\text{Sustained}|\text{Success})$

Figure 6-3: Mechanisms to Support Individual Decision Making

### 6.2.1 Minimizing Decisions

Many of the above decisions are difficult for the untrained person to make effectively. Some of these decisions can be made automatically, based on prior experimental outcomes; others can be partitioned into a design problem that can be amortized over many user's experiments. By properly structuring the interactions a user has with the system, the minimal decisions that the average user needs to make may be reduced to:

1. What symptom do I want to improve?
2. What intervention is **most likely** to work, and has **acceptable burden**?
3. What measurement capabilities do I have available?

Alice's decision is reduced from a complex technical design problem to a filtering and selection problem. Her symptoms or diagnoses can be used to filter and sort a list of matching interventions and associated experiments. The sorting of the results can be based on the probability of success, along with documentation of the costs of the trial. For example, Alice saw that glycerin was "low hanging fruit," being both easy to test and likely to work. The dietary treatments were harder to test and less likely to work, and she was able to judge what was most appropriate for her at each point in time. During trial creation, the user can choose to use devices that are more precise and reduce the expected time needed for the trial based on what they have available. The mechanisms required to support the decisions are summarized in Figure 6-3.

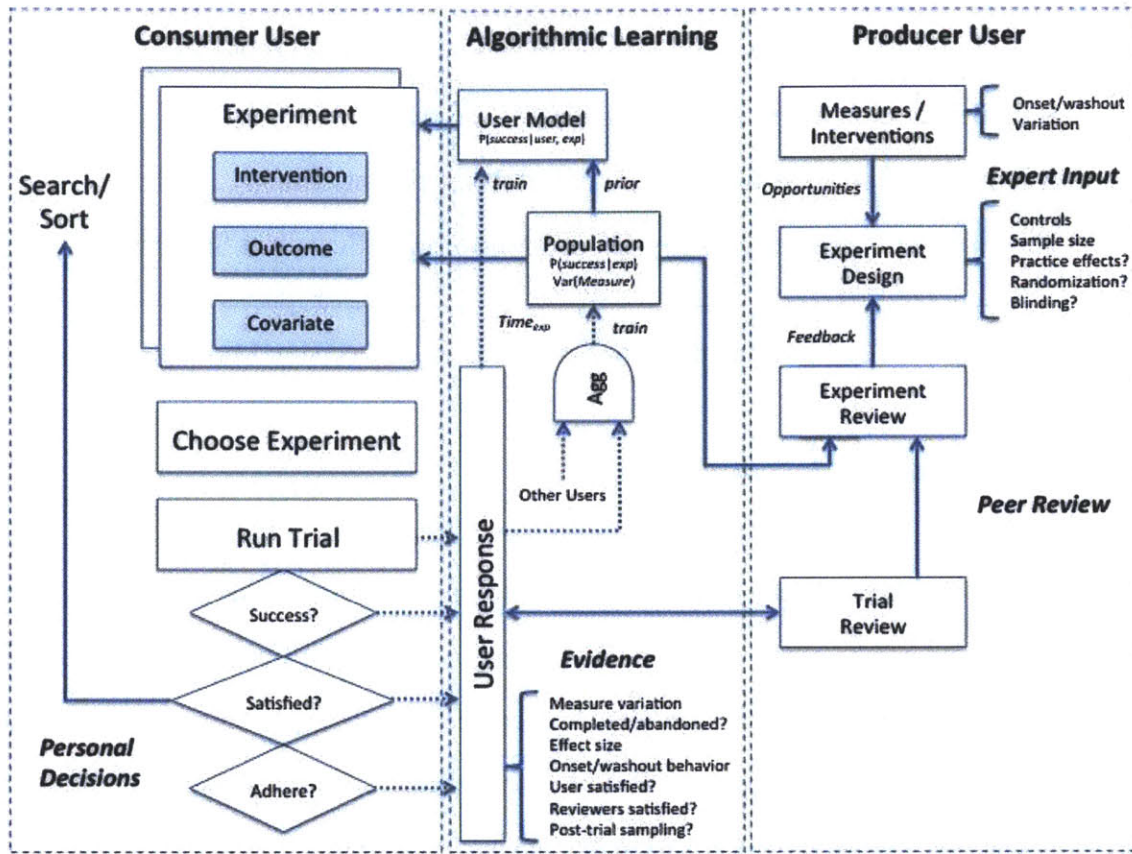


Figure 6-4: Process Model

The value of the simplified experiment is a dramatic simplification of an average user's choices, while retaining the ability for more ambitious, skilled users to engage with the full complexity of the experiment design problem by defining new experiments and reviewing trial results. Individual trials produce empirical data that both informs the expert's experimental design and optimizes the individual's trials. The written experience of others who had successful trials can inform the patient's personal decision about experimenting with or continuing an intervention. Figure 6-4 is a process diagram that illustrates information flows among these three processes of trial execution, automated learning, and experiment design.

Once selected, running an experiment consists of combining one or more tracking tasks with the application of an intervention on a specific schedule. In addition, an action plan is required to accommodate unanticipated events that may violate the assumptions of an

- **Sharing** A user can share their data anonymously, with specific individuals, or open it for public comment. This improves individual interpretation, contributes to cultural learning, and enables algorithmic analysis of trial outcomes.
- **Peer Review** Sharing enables feedback, allowing other users to discuss the trial or experiment with each other and to rank whether the experiment design, individual trials, and results of aggregation are well conceived.
- **Replication** Each user can replicate their own response for increased internal validity, and multiple users running trials helps establish the external validity of an experiment.

Figure 6-5: Community Science

experiment (e.g. a stable baseline, good adherence, etc).

### 6.3 Designing Experiments

The framework assumes that sufficient methodological expertise exists in any given group to formulate experiments, characterize measures, and document treatments. These *producers* generate new hypotheses for the other users to consume. The *consumer* users test the hypotheses and generate feedback for the producers so they can generate improved hypotheses for evaluation. However, the design of an experiment is a challenging task and one that professional scientists get wrong every day. For example, it is easy to underestimate population variance leading to under-powered studies with indeterminate results. How can we support ambitious, but untrained, users while minimizing time-wasting methodological errors that are routinely made by trained professionals?

The solution to this problem is the same as that used by professional scientists. Introducing structured processes of feedback helps to ensure that errors are identified and fixed over time, bending the long-term behavior of the system towards accuracy. Figure 6-5 lists the additional steps of a scientific method that apply to the Aggregated Self-Experiments framework.

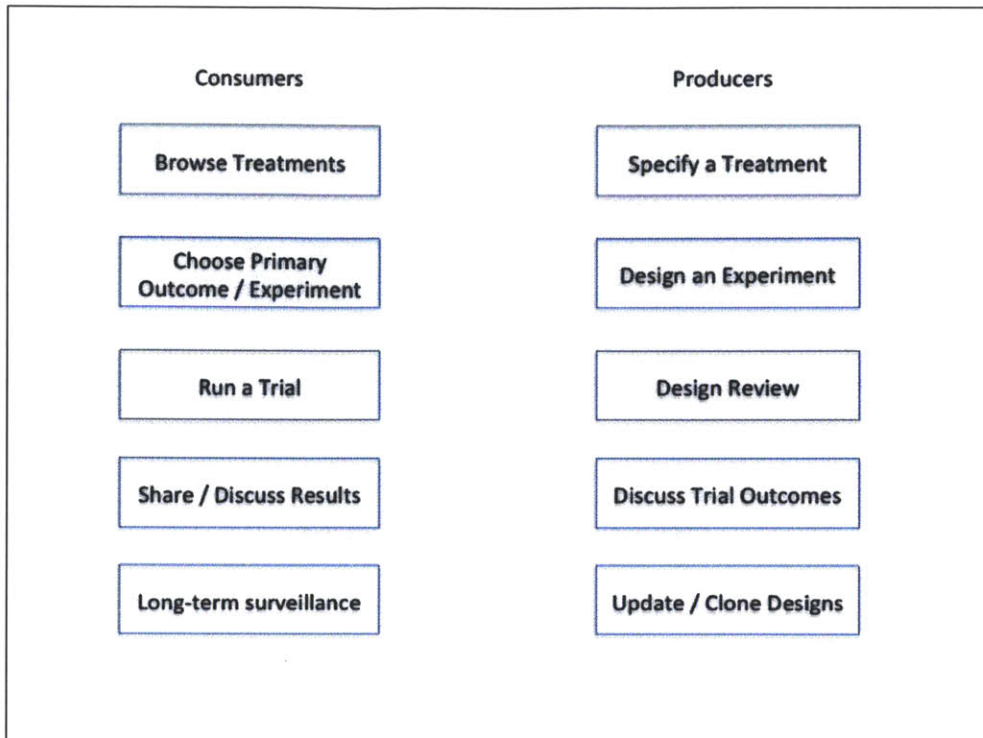


Figure 6-6: Producers and Consumers

The ratio of content producers to consumers in online communities is highly skewed towards the consumer. Pew reported that less than 10% of online users seeking health information actually contributed back information of their own [FD13a]. For consumers, tasks must be made as approachable and relevant to their goals as possible, whereas tools for producers must enable them to address the inherent complexity of the design task. The framework must ensure that the activities of consumers are well leveraged so their time is well spent, without requiring a significant amount of direct interaction. Figure 6-6 shows the actions needed for each class of user in supporting the process diagram of Figure 6-4. Any user can act in one or both capacities, but the design goals and expectations on tools for each role may be very different.

### 6.3.1 Sharing and Peer Review

Peer review is supported by allowing users to share their individual trials, and by allowing any user to review aggregate population data produced for a given experiment. The mech-

anisms for feedback include ranking (“this is a bad design”), commenting (“this is what I learned”), and the history of choices users make to run a trial and continue the treatment afterwards.

### **6.3.2 Replication**

Replication for a set of n-of-1 trials has different implications than replications of population trials (Section 2.5). In Aggregated Self-Experiments there are two purposes to the replication of an experiment when run within or between patients.

The first purpose of replication is to increase a user’s confidence whether a treatment works for them. This re-test may be triggered by a change of condition, a return of the original symptoms while on the treatment, or a desire to demonstrate the response to others.

If the trial is a success, the user will want to attempt to maintain that treatment over a longer time period. The framework can sample how well the intention is carried out by engaging in ongoing tracking (“surveillance”) or periodic follow-up (“sampling”), to see whether the user is still engaged with the treatment and whether the outcome improvement has been sustained.

If the user is no longer engaged with the intervention, or is not seeing the benefit identified during the trial, there are several possible interpretations:

- Adherence failure. The user was unable to sustain the treatment.
- A false positive. The trial was positive, but for the wrong reason.
  1. Placebo response
  2. Confirmation bias
  3. Unmeasured confounder
  4. Flawed trial design
- A change in user status. The user’s health baseline has changed.

The second form of replication involves trials run by multiple users. These replications enable estimation of the dropout rate, prevalence of successful trials, and the average effect

size across successful trials. The more users who run an experiment, the more accurate these estimates. Multiple trials also provide feedback on flaws in the design, such as treatments with longer onset or carry over than expected, the protocol, or the side effects of the intervention. Section 7.5 details specific mechanisms for computing these factors.

### **6.3.3 Improving Designs**

Producers can update experiments based on feedback from both population outcomes and explicit feedback answering questions such as:

- Is a noisy measure predicted by a more precise measure? (Suggests using an alternative outcome)
- What is the observed transition period? (Suggests changes in onset or washout periods)
- Reported side effects (Update treatment documentation)
- Previously unknown or unmeasured confounders (Adapt the design)
- Unexpected observations (Special cause can yield new hypotheses)

## **6.4 Reducing Trial Burden**

Adherence to an experimental protocol, even when motivated, is difficult for many people. Reducing the time of an experiment and increasing its robustness to the noise of everyday life is crucial. As use of the framework scales, there will be a growing corpus of measurement time-series and experimental outcomes. As discussed in Section 2.5 and Chapter 7, the time taken to run a trial is determined by the number of phases, onset and washout periods, and the sample size per phase. The sample size within a phase is a function of:

- Desired minimal effect size



- Desired confidence in the trial result
- Prior belief of effect size
- Type of the outcome measure
- Variance of the outcome measure (inherent + measurement error)
- Autocorrelation of the outcome measure
- Frequency and magnitude of reported or measured confounding factors

The first two factors are determined by convention or individual choice, but the remainder can all be estimated from evidence. The computation of sample size is the subject of section 7.2.3 and the estimates of the parameters from individual and population data is discussed in section 7.5.

### **6.4.1 Optimizing Sample Size**

Without information from the population the trial design has to make conservative assumptions about variance, or it risks under-powering the trial to detect a given effect size. A well-characterized measurement enables informed, often much tighter, assumptions about variance during trial planning and can adapt to surprises in observed variance as the trial commences. The net effect is to reduce the number of planned samples per period (see Section 9.3.2). Another common assumption in statistical analysis of any trial, as well as N-of-1 trials, is that the variance in both arms of the trial is identical. This may or may not be true for the measures here. Experience with the trial platform suggests that some treatments have much lower variance under treatment than baseline, which provides another opportunity to tighten the trial schedule. Reduced variance may itself be an endpoint of interest or an enabler to future experimentation (by increasing the sensitivity of the outcome measure).

After a number of successful trials have been executed, the framework can estimate the inflection point at which the effect mean stabilizes, improving the accuracy of the onset and washout adjustments which may shorten or lengthen the overall trial schedule. For many

treatments, this is the most inflexible and time-intensive part of estimation, and prospects to exploit, rather than ignore this period during each phase are discussed in Section 10.1.3.

Chapter 7 discusses the Bayesian machinery used to interpret data in this framework. One of the critical elements of a Bayesian interpretation is the selection of a prior. Without population data, it is prudent to select an uninformed prior distribution over the effect size. With population data it is possible to have an empirically determined prior. Assuming the effect size is consistent across the population, the fit and confidence of the prior increases as more patients run the trial. An accurate prior reduces the data required for a given user's trial to achieve confidence in observing a treatment effect. Section 9.3 reports the magnitude of improvement possible with accurate priors.

## **6.4.2 Accommodating Confounding Factors**

The above analysis largely ignores the challenges imposed by confounding influences. The other trial design parameter that contributes to long trials is the choice of design. Trial designs can be categorized into the categories visualized by Figure 6-7.

The time-series properties of the measured outcome and covariates, when correlated to the treatment period, provides strong signals regarding the outcome response to the intervention. Treatments with strong effects not typically subject to confounding may be adequately tested by simple interrupted time series designs. By contrast, treatments with weak effects, high variability, and many known confounders should be subject to a number of crossovers. For outcomes with a few, limited and highly reliable confounding factors, it may be possible to adjust the outcome measure to account for those influences.

This approach prefers characterization of confounding influences over extensive crossovers to control for them. This is a necessary step in dealing with self-experimentation because of the difficulty of ensuring adequate controls. A characterization strategy is much more likely to be successful under this framework because of the potential for achieving an empirical scale well beyond anything produced by existing clinical or academic research. Data

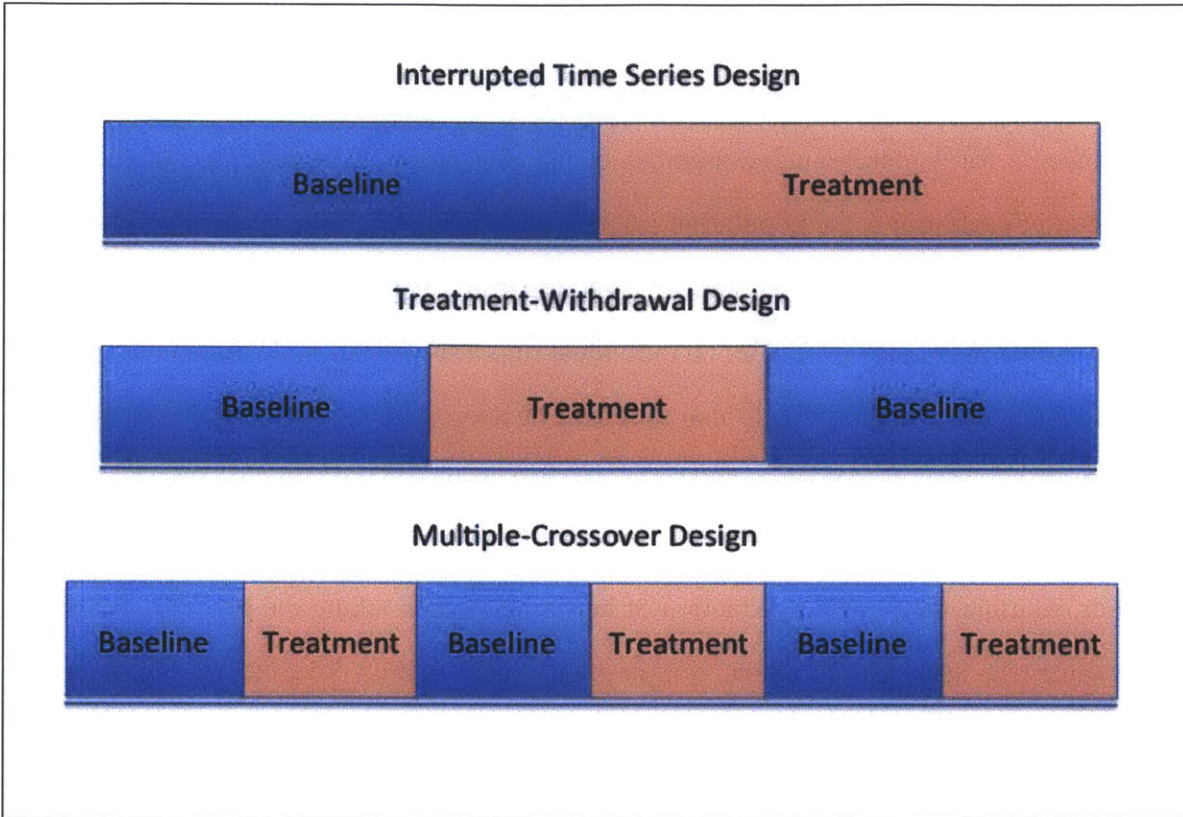


Figure 6-7: Common Single-Subject Designs

sharing under this framework is far easier and more seamless than in most clinical trial and academic research settings (e.g. see [ES10]).

## 6.5 Computing the Probability of Success

The most expensive of all user activities is testing a treatment that has no effect and that the user would not otherwise have tried.<sup>3</sup> Being able to sort experiments by their probability of success is a key goal of the framework and this can be most simply computed via an estimate of the base rate of trial success over all users.

Deciding on the efficacy of a treatment is nuanced and addressed in discussion of prior work in Section 2.5 the second half of Chapter 7. Setting aside philosophical debate about generalization from samples of one, we can treat repeated trials of an experiment as draws

<sup>3</sup>A negative result is valuable if it helps a user discard a treatment they believed, a-priori, was effective.

from a Bernoulli distribution.<sup>4</sup> The sample mean of binary outcomes is the maximum likelihood estimate of the parameter of the distribution, or the probability the next trial will be a success for a representative member of the population. The posterior of the belief update after each trial is the population prior  $\text{Beta}(\alpha_{ijk}, \beta_{ijk})$ .

In the absence of other information, each user can simply use the population prior as their personal prior. However, after an individual runs a successful trial with an intervention for one measure, it is more likely that they will be successful on a different measure, particularly if we know that the two measures co-vary or the second outcome was a covariate of the first trial.

For example, Alice saw a reduction of her itching when taking the glycerin treatment. That same trial also showed a reduction in the psoriasis scaling covariate. If she ran a trial of glycerin for scaling, she is much more likely to get a positive result after the first trial than any member of the population. Similarly, if her fatigue responded to one restrictive diet, it is more likely to respond to a related diet.<sup>5</sup>

## 6.5.1 Recommending Experiments

When enough people have run a trial of an experiment, the probability of the next person succeeding (assuming users are exchangeable) is well-represented by the Beta distribution induced by the empirical distribution of outcomes. That is to say, given a set of experiments, I can choose the experiment for which the prior probability of success is the highest. Using the probability of success alone is unlikely to be sufficient in practice, because there are other parameters besides the binary success that come into play.

- **Probability of Effect** How likely is this user to see a given magnitude of effect?

---

<sup>4</sup>Where the distribution models the self-selected population of people who come to the website and run an experiment

<sup>5</sup>While the prototype built for this dissertation does not implement these predictions due to the lack of a pre-existing corpus for analysis, using trial outcome as a predictor of other trial outcomes may be a powerful way to personalize recommendations.

- **Magnitude of Effect** How large is the effect of the treatment?
- **Preference** Will the user be satisfied with a treatment if it works?
- **Long term success** Will the user continue this treatment long term?
- **Learning** Will the community as a whole benefit from additional information about the treatment?

## 6.6 Deciding to Continue

The decision to maintain an intervention is based on more than whether the experimental outcome was positive. For example, users learn more about the true costs of the intervention during the trial. Can I really sustain this? Is the effect size large enough to justify the investment? Can I afford it? Perhaps the most important question is: if I stick with it, how likely is it that the positive result I saw in this trial will be sustained over the coming months and years? After the trial, the judgement of the cost of the intervention will change and that influences a user's desire to continue, regardless of the trial outcome.

Further, a successful trial does not mean that the intervention itself is effective, nor that the user is guaranteed to get the same results the next time. There will always be confounding factors that could provide an alternative explanation for the trial outcome in preference to a treatment response. A well designed, well-run trial will increase confidence in the intervention's efficacy, but the a successful trial outcome does not necessarily imply that the intervention will be useful over the long term. This discussion is continued in Chapter 7.

## 6.7 The Population Hypothesis Space

There are refinements to the process model to address scientific processes that go beyond the support of individual decision making emphasized here. There are questions that may

be of interest to a community at large that motivate individuals to contribute to population level knowledge.

### **6.7.1 Predictor Discovery**

Most treatments, if effective, yield a highly heterogeneous response; few treatments help everyone.<sup>6</sup> Most treatments act differently in different people, and have widely varying side-effect profiles. Aggregation of outcomes allows us to estimate the prevalence of successful trials as well as a population effect size.<sup>7</sup> Ideally we would like to figure out for whom the treatment will actually work, and for whom it will cause untenable side-effects.

Most people are not testing interventions from an untreated baseline. The baseline is already a personal standard of care with the new treatment added / withdrawn. This will have an effect on the prediction for the magnitude of effect in that person as well as across the population. Documenting a user's current standard of care, along with other background information, can provide a feature set that may more accurately predict individual trial success and/or effect size.

Anyone in the community can introduce a question given to each user of an experiment or intervention that they feel may identify a sub-population that predicts side effects or successful vs. unsuccessful outcomes. Observed associations can be used to modify the conditional probability of a success or the effect size. Predictors help address the first optimization problem introduced above: helping consumers select an experiment to run.

### **6.7.2 Causal Analysis**

Patients will often test combinations of interventions (e.g. see A.2.1). Their symptoms will improve, but we don't know to what specific intervention, or subset of interventions, to

---

<sup>6</sup>Even sleep, healthy eating, and exercise, though commonly effective for a wide variety of ailments, are detrimental to people with specific genetic or infectious conditions or at some points in time, illustrating that universal applicability is rare.

<sup>7</sup>These values only have internal validity for other people acting within the framework, a subject touched on further in Chapter 7

assign as the causal agent. From the standpoint of this framework, it is irrelevant so long as the combined intervention has a reliable effect. However, combination therapies can be onerous or expensive to carry out, reducing adherence and successful replication. If some or most components of the treatment are superfluous, then there is room for narrowing the hypothesis by creating one or more sub-treatments and experiments to go along with them. Producers can encourage users who succeeded with a combination therapy to try a sub-therapy to see if they get the same result.

### **6.7.3 Ethical Considerations**

Enabling non-scientist self-experimentation may give many medical professionals cause for concern [Mar08]. An easy, but perhaps unsatisfying answer is that this activity already takes place at great scale, as discussed in the introductory paragraphs of Chapter 1. This framework simply provides tools that seek to reduce the error taking place in these settings. However, new tools also make possible new kinds of unforeseen mistakes.

For example, the “spreadsheet” disease is a potential future consequence [Pan98]. People are comfortable with the uncertainty inherent in their own memory, but when we delegate reasoning to a tool, there is a strong inclination to take answers at face value. It often requires training to understand how to diagnose problems that are “garbage-in, garbage-out.”

Further, using these tools for population-level research will often be the result of a central individual or group encouraging others to try a treatment. This raises the ethical (and possibly legal) obligation that should or will be placed on these users. In this new model, what is the equivalent of human subjects review? Who ensures that recommendations are fair? How do we ensure that people are not being exploited by commercial interests?

The approach I recommend for optimizing for ethical conduct parallels the ethical framework of Elwyn et. al. [EFT<sup>+</sup>12] and the authors of a recent monograph on Learning Health Systems [FKG<sup>+</sup>13]. The algorithms used to make recommendations should

be public and reviewable. Producers should be personally identified. Data used to make recommendations should be reviewable in its de-identified form. Transparency and crowd peer review is the best prescription to avoid ethical problems caused by new kinds of tools.

Allowing users to comment on treatment side effects or suspicions about an experimental protocol will be critical to ensuring that a future user is maximally informed about the risks and potential rewards of a treatment they are considering. This is an improvement on today's ad-hoc strategy of reading various reports on forums to establish an assessment and probably superior to most informed consent documents [LLS11]. Reputation systems can be used to adjust for the credibility of the report [JIB07].

Finally, to the professional it is important to remember that the interventions addressed by this framework are exactly the kind that are unlikely to fall under any significant clinical review. When they do, patients should consult with a professional. But the treatments that the medical profession is largely ignoring - behavior, psycho-social, supplements, exercise, etc. - carry little danger and are already tried out by patients without supervision in large numbers. For example, under no circumstances should population parameters such as mortality rate or medical complication rate be an outcome variable of interest. Errors instead represent opportunity costs for users because they are spending their time and money on something they shouldn't, or bypassing an intervention that could help them. If errors are a first class consideration of the framework, expected by users, and transparently reported – then we can improve this framework by identifying and recovering from errors as they occur.



# Chapter 7

## Algorithms

The prior chapter identifies a concrete set of questions faced by users in the Aggregated Self-Experiments model that may be assisted by an analysis collected data. These are :

- How long do I run the trial phases?
- How many phases?
- Was the trial successful?
- Should I continue the treatment?
- What is the best trial for me to try next?

The chapter also identifies ways in which aggregate data can be used to improve the accuracy or efficiency with which the system answers these questions. Accordingly, this chapter introduces a set of simple algorithms to support individual decision making and system optimization.

### 7.1 The Single-Subject Trial as Decision Aid

The purpose of the Aggregate Self-Experiments framework is to help individuals make better decisions about their own care by improving the quality of the information they have to work with. Informally, we want to answer the question “should I take/continue this intervention?” Decisions such as this can be discussed more precisely using the language

of probabilistic decision theory. Decision theory characterizes a decision problem  $D$  as the maximum of the expected utility  $EU$  of one or more actions  $A$  over a set of outcome values  $O$  given a utility function  $U : [O] \rightarrow \mathbb{R}$ , a cost function  $C : [A] \rightarrow \mathbb{R}$ , and an outcome model  $\pi$  that maps actions to a probability distribution over outcomes [KF09]:

$$EU[D[a]] = \sum_{o \in O} \pi_a(o)[U(o) - C(a)] \quad (7.1)$$

The outcome in question is the direct effect of the intervention on the trial's primary outcome,  $\delta$ . For simplicity, outcomes can be a continuous function of the effect size, or broken into five discrete levels: negative, none, small, medium, and large. The cost function represents the burden imposed by trials and treatments with regards to time, money, mental energy, or potential side effects. The actions Alice can take in this context are:

- Stop the intervention / do nothing:  $a_0$
- Adopt / continue with the intervention:  $a_1$
- Perform a trial of the intervention:  $a_e$

If we have high confidence the treatment works, there is no need to experiment. If we have very low confidence, then there is no point in continuing a treatment. Experimenting is helpful when Alice is unsure about the costs and benefits of the intervention and wants to make an informed decision between  $a_0$  and  $a_1$ . As discussed in section ??, the Bayesian analysis of trial data affords a wide range of useful inferences. To make a decision to sustain an intervention, the utility of the intervention effect weighed by its probability minus its cost should be greater than the utility of no intervention. Given a preliminary utility and cost function, it is possible to solve for the necessary level of confidence  $\eta$  we have that an intervention will yield an expected effect size of  $\delta$ , motivating a user to chose the treatment.

$$P(\Delta\mu = \delta) \geq \eta, 0 < \eta < 1 \quad (7.2)$$

Performing a trial of a treatment with which the user has no experience also changes the user’s perception of utility and cost. A trial effectively updates the user’s probability distribution over effect size  $\pi$ , the subjective utility of the effect  $U(o)$ , and the subjective cost of the treatment  $C(a)$ . For example, the cost of the dietary treatment was initially too high for Alice, but when she found a shorter treatment that would raise her confidence, she decided it was worthwhile to learn more. After she experienced the benefits of the diet, she decided that the cost was more than worth the value.

In practice, identifying concrete utility and costs functions is challenging (??). Moreover, real utility functions change with time and circumstances. An approximation is to ask the user how large a change they want to see to continue the treatment  $\delta$ , and how confident they want to be in the result  $\eta$ . This probes the utility and cost function indirectly without requiring that an exact function be determined.

When  $\delta$  is small and the variation of baseline measure  $\sigma_0$  is high, it is possible that an observed  $\delta$  meeting equation 7.1 is due to ordinary variation and not a treatment effect. Using Bayesian estimation techniques [Kru13] we can estimate the probability distribution of the effect size. The decision rule above can be rewritten to state that  $\delta = 0$  lies outside the a one-sided highest density interval (HDI) of width  $\eta\%$  over the credible values of the estimate.

$$0 \in \text{HDI}_\eta(\Delta\mu) \tag{7.3}$$

This is analogous to rejecting the null hypothesis at  $\alpha = 1 - \eta$  in traditional null-hypothesis testing.

To run an efficient trial of a treatment, Alice will need to tell us the minimal size of effect  $\delta$  she is willing to look for and how confident she wants to be in any observed outcomes  $\eta$ . The system will then parameterize a concrete trial to accumulate data whether condition  $T$  changes the expected value of a symptom  $O$  by at least  $\delta$  with probability greater than  $\eta$ .

After the trial has been run, and the system's belief about the treatment effect updated the system can check the actual power of the conclusion and add additional baseline/treatment pairs to accumulate more evidence if the current belief neither finds a treatment effect or has confidence in a null effect.

## 7.2 The Ideal Trial

A self-experiment to evaluate treatment  $T$  will compare a baseline measurement of the outcome symptom  $O$  without  $T$  to a measure of that symptom with  $T$ . The value of  $O$  is determined using a specific measurement  $M$ . The following treatment represents the simplest of possible designs and analysis, following the A-B, or interrupted-time-series design (??).

### 7.2.1 Measurement and Variability

Most measures  $M$  in a self-experiment (such as mood, weight, etc) will vary from sample to sample throughout a given baseline or treatment phase. Nearly all measurements imperfectly capture the true value of the underlying phenomenon, modeled by adding an error  $e_m$  to each measurement. Error terms may consist of either a static bias and/or a random effect. This combination of inherent variability and instrument-determined variability results in a measurement variability that must be accounted for in comparing baseline data to treatment data.

For example, if a trial tests a hypothesis that melatonin increases the amount of deep sleep and deep sleep varies +/- 5 minutes every night and the treatment improves it by 10 minutes, then a single high measure during baseline compared to a low measure during treatment would appear to show no treatment effect when in fact there was an increase. If the measures are reversed, but the treatment has no effect, we would see a 10 minute spread and assume there is a treatment effect of size 10. The solution to this problem is to take

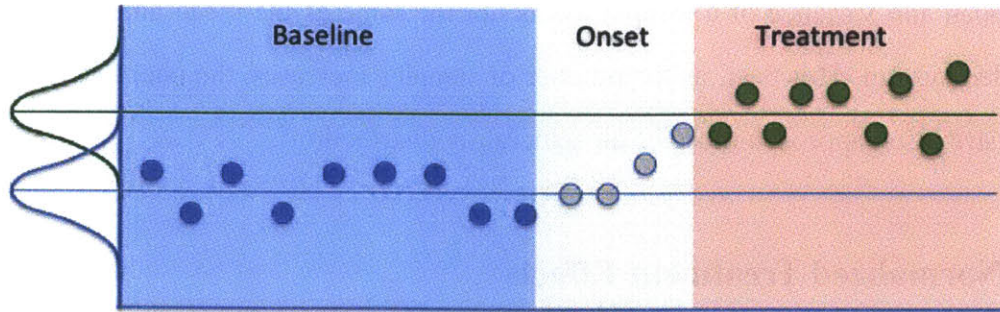


Figure 7-1: AB Trial with Treatment Effect

repeated measurements in each phase of the trial and compare a statistic computed from each measurement series.

Processes that fit the normal distribution  $N(\mu, \sigma^2)$ , a reasonable approximation for many measures, are parameterized by the statistics mean  $\mu$  and variance  $\sigma^2$  or standard deviation  $\sigma$ . The square root of the variance is called the standard deviation  $\sigma = \sqrt{\sigma^2}$ . The interpretation of these statistics is straightforward. 68% of all observed values are expected to fall within one  $\sigma$  of  $\mu$  and 95% of all values are found in the range  $\mu \pm 2\sigma$ . The variance of the model characterizes both the intrinsic variation of the measured process and uncertainty introduced by sampling<sup>1</sup>.

The measure  $M$  defines a procedure or a physical instrument for sampling values at multiple points in time. These sample measurements  $M_0$  and  $M_T$  can be used to estimate the parameters of a model of the underlying process in each phase. Figure 7-1 provides a visualization of the generating model for data collected during two trial phases when the intervention changes the underlying process. The figure also introduces the notion of an onset time during which the process is in transition between the two models and the values of the measure are ignored.

<sup>1</sup>This holds under the assumption that the mean of the process is stable. Unfortunately, many measures are not stable over time. A sustained rate of weight loss (or gain), introduces an overall slope in the observed data. The model developed here emphasizes models that are locally stable, meaning the parameters of the model are assumed to be fixed through the duration of the trial. This is a common situation for many chronic diseases and lifestyle factors like productivity and sleep. Ongoing surveillance of a user can be used to detect post-trial changes that might motivate a re-evaluation (??)

The mean and variance of a sample  $M_0$  is not the same as the mean and variance of the modeled system. However, as the number of samples increases, the parameters of the sample distribution approach those of the generating distribution<sup>2</sup>.

## 7.2.2 Normalized Treatment Effects

The effectiveness of a treatment is assessed by taking the difference between the two sample means  $\delta = \mu_T - \mu_0$ . The magnitude of measures vary from satisfaction scales (0-7 numbers) to large real value ranges (e.g. blood pressure), so the value of a minimally meaningful effect is different for every measure. Cohen [Coh88] introduced a standardized effect measure that normalizes by the observed standard deviation:

$$d = \frac{\mu_T - \mu_0}{\sigma} \quad (7.4)$$

Cohen also analyzed a wide variety of psychological experiments and suggested a subdivision of standardized effect sizes into categories that roughly correlate to clinical importance:

- **Small**  $d < 0.25$
- **Medium**  $0.25 < d < 0.8$
- **Large**  $d > 0.8$

Figure 7-2 shows three examples of Cohen's  $d$  for different generating distributions. While this division is not universally applicable, and has proven to be somewhat controversial, it provides a practical starting point for standardizing statistical power targets across a wide range of experiments.

If an effect is large relative to the measured variability in each phase, then the two datasets are clearly separated and a different set of parameters must have generated the two

---

<sup>2</sup>Assuming no additional confounding factors are present. This assumption will also be relaxed in future sections

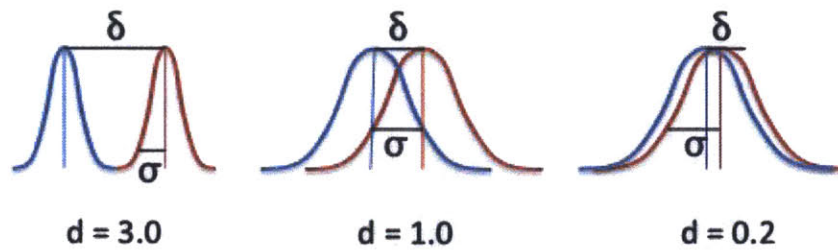


Figure 7-2: Cohen's  $d$  Effect Sizes

sets of observed data (figure 7-2,  $d=3.0$ ). If, however, the variability is larger than the effect size, then the two models are nearly identical (i.e. their means are close together) and it requires a far larger number of samples to be confident in a significant difference between the means. Thus a key part of trial planning is determining how much evidence we need to detect a change of a given size, and the analysis of a trial requires assessing whether the evidence we received is sufficient to predict a treatment effect.

### 7.2.3 Planning a Trial

Each trial takes place with a user-provided, or system default, assumptions about the minimal normalized effect size  $d$  and the minimal confidence we want in the outcome if an effect of that size were actually observed. Trial planning involves preparing a schedule of baseline and intervention phases according to the experiment's specification. The baseline phases and treatment phases need to have sufficient samples to power the trial to detect  $d$  with confidence  $\eta$ .

To assess the sample size given these parameters I employ the model outlined by Kruschke [Kru13] which uses simulated draws from an ideal distribution to compute the sample size at which  $\eta$  percent of trials with an actual effect of size  $d$  conclude that there was a treatment effect.

Figure 7-3 illustrates a model of two t-distributions with parameters of mean  $\mu$ , standard deviation  $\sigma$ , and spread  $\nu$ . The density interval  $\eta$  is the interval of highest density (HDI) over the posterior parameters; a trial is successful if the mean of the baseline model is not

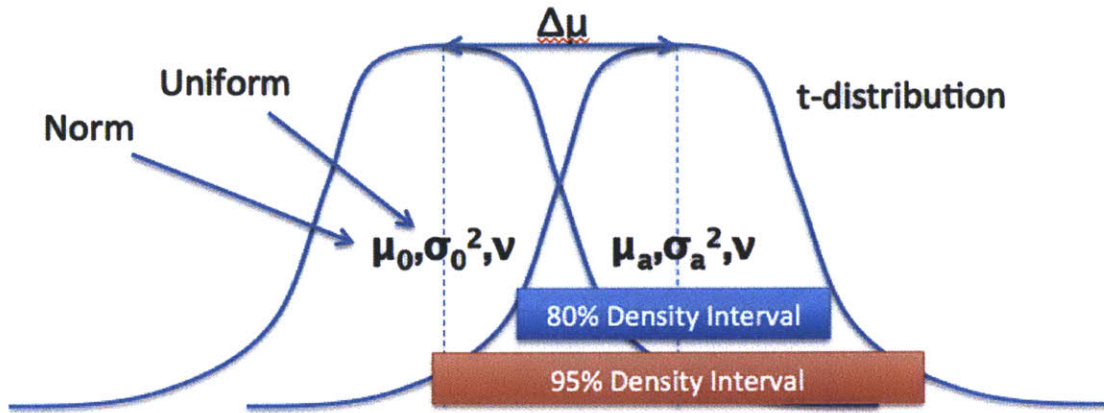


Figure 7-3: Statistical Model

contained in the HDI of the posterior treatment parameter distribution.

The model assumes priors of the form:

$$\mu = N(\mu_p, \sigma_p^2) \quad (7.5)$$

$$\sigma = \text{Unif}(L, H) \quad (7.6)$$

$$\nu = f(\lambda) \quad (7.7)$$

Where  $f(\lambda)$  is a shifted exponential for  $\nu > 1$  and  $\lambda = 29$  according to the method of Kruschke. This model is robust to outliers in the source data. Computation of power is performed by repeated computations of the posterior for different parameters induced over multiple runs given a fixed set of samples drawn from an ideal distribution with some percentage of outliers added.

In contrast to more traditional trials, each user's sample size can be customized to prior data about the user, outcome measure, and prior experimental outcomes (see below and Section 9.3). In the case of a trial with no prior information, we estimate  $n$  using a conservative model with uninformed priors over both mean and variance as above.



## 7.2.4 Updating Beliefs

After a trial is run, the measurements observed during the trial are used to incrementally update the system's belief about the behavior of the patient during baseline and treatment phases; the difference of the two is the treatment effect, if any.

$$\pi_a(o) = P(o|\boldsymbol{\theta}, a) \quad (7.8)$$

Here,  $o$  is a random variable representing outcomes in terms of Cohen's effect size.  $\boldsymbol{\theta}$  represents one or more formal parameters of a model that parameterizes a distribution over  $o$ . Because doing nothing, sticking with the baseline, guarantees an effect size of 0 under assumptions of stability, we only need to update the belief about  $\pi_1(o)$ , the probability of a meaningful effect if Alice chooses the intervention.

One procedure for updating model parameters based on observed evidence is Bayesian belief updating. Bayesian belief updating dictates that the posterior distribution of the parameters is the normalized product of a prior belief and likelihood of the observed evidence:

$$P(\boldsymbol{\theta}|\bar{x}) = \frac{1}{Z} \mathbb{L}(\boldsymbol{\theta}|\bar{x}_a) P(\boldsymbol{\theta}) \quad (7.9)$$

$\bar{x}$  represents sampled evidence. The likelihood function  $\mathbb{L}(\boldsymbol{\theta}|\bar{x})$  characterizes a distribution over parameters and is equal to the probability of the observed data given the model parameters. The outcome distribution is the likelihood multiplied by our prior beliefs about the model  $P(\boldsymbol{\theta})$ . The design of the trial dictates the association between collected data and a given action  $a$ , such that  $M_0 = \bar{x}_0$  and  $M_T = \bar{x}_1$ .

Under the Bayesian model, if we get a null result that is sufficiently powered, then we can conclude the treatment is likely to be ineffective. If however, we find that the

trial wasn't sufficiently powered, we can add an additional treatment-baseline pairing<sup>3</sup> to accumulate more evidence until our model is sufficiently powered to satisfy our confidence targets.

The model of Figure 7-3 can be given empirical priors learned from this belief updating procedure.

## 7.2.5 Interpreting the Trial Outcome

The overall success of a trial is determined by running a belief update procedure on the trial model against the observed data. If the posterior treatment distribution's HDI does not include the mean parameter of the baseline distribution, then the trial is deemed a success.

## 7.2.6 Visual Outcomes

Visual analysis of the data is indicated as a crucial capability by the literature. We chose to use a variation of the Shewart control chart (see section 8.4.2 for a detailed description) to plot measurements against a time axis, overlaying the phases of the experimental as a shaded regions. Upper and lower control limits are computed as guides to significance. Points that are above or below the control limits during treatment periods are very likely to be drawn from an alternative distribution to that observed during baseline. Data points within the limits that represent shifts in the mean measure (such as runs of points above or below the mean) are also highlighted as evidence of a treatment effect.

Typically, Shewart's control limits are computed at the three-sigma significance level, representing a 98.6% probability that there is a significant shift in the mean of the process. This was chosen to avoid false alarms when monitoring factory operations; at 2-sigma confidence an alarm would be found on average every 20 days. In Aggregated Self-Experiments, there is a direct mapping from the decision criterion to the control limit

---

<sup>3</sup>Under the assumption that exchangeability holds over short timescales

calculation, allowing the preferences expressed in the trial design to be visually apparent in the decision rules and subsequently rendered control charts.

The control limits are chosen based on the treatment sample size  $n/2$ , the confidence  $\eta$ , in units of  $\sigma$  from the mean  $\mu$ . Namely, we choose our limits such that the probability of seeing a value  $k$  standard deviations outside the limits in  $n/2$  samples purely by chance is  $\eta$  times the number of samples (equation 7.12).  $k$  can be estimated directly from the cumulative distribution function.

$$\text{ucl} = \mu + k\sigma \quad (7.10)$$

$$\text{lcl} = \mu - k\sigma \quad (7.11)$$

$$\frac{n\eta}{2} = \frac{1}{(1 - \Phi(k\sigma))^{\frac{n}{2}}} \quad (7.12)$$

$$\frac{2}{n\eta} = (1 - \Phi(k\sigma))^{\frac{n}{2}} \quad (7.13)$$

The CDF for a t-distribution should be used if the baseline measure only has a small number of samples so far ( $< 40$ ). Additional control rules for sequences within the control limit can be computed similarly, although including more rules increases the overall false alarm rate so. The use of multiple rules should be adjusted; dividing the probability of a false alarm by 4 for any given rule is a reasonable heuristic [CW87].

### 7.3 Working with Real-world Data

The above model assumes ideal conditions, namely that each sample is independently and identically distributed (IID), there are no other influences on the outcome variable, and the user adheres perfectly to the protocol. None of these simplifying assumptions hold in real world settings.

### 7.3.1 Time series effects

The most common violation of the IID property is one where the current measurement systematically depends on one or more prior measures. For example, if a user has a bad night of sleep, it effects the next several days, with declining effects each day. This is called autocorrelation.

The magnitude and number of samples of dependency of an autocorrelated measure is related to the sample rate. For example, if I measure my mood every minute, my mood now is very well predicted by my mood. However my mood today may have little to do with my mood tomorrow, or my mood may only vary on longer time frames, meaning it takes many days to transition from one state to another.

A second time series effect of interest is typically referred to as seasonality. Of course in our self-experiments, we rarely will see experiments that have seasonal dependencies but we do see many experiments that have periodic dependencies on the day of week or time of day. Time of day dependency can be caused by fatigue that sets in later in the day, leading to consistently less productive afternoons than mornings. User behavior during weekdays is also different than weekends. Users on weekends may engage in more exercise, get more sleep, and have much less consistent schedules. We adjust for weekday effects in the same manner as seasonality, by training a model on prior observations to adjust the observed signal for the seasonal effect, or by ignoring samples collected on the weekend (as in online productivity, were most work happens during the week).

The solution to both these problems is to characterize the time-dependent influences on each measure and subtract them out to get a more accurate estimate of true value of the measured variable. Unfortunately, both of these adjustments remove redundant information encoded in the signal and require somewhat higher sample size to compensate.

There are standard models for accommodating autocorrelation and seasonality. Prior work and a preliminary analysis of data collected in this study indicates that a single day lag model accounts for most of the observed correlations.

The AR(1) model is defined as:

$$x_t = \mu + \phi(x_{t-1} - \mu) + \epsilon_t \quad (7.14)$$

Where  $\mu$  is the mean of the series,  $\phi$  is the amount of influence from  $x_{t-1}$  to use to predict  $x_t$  subject to a random error term  $\epsilon_t$  typically assumed to be normal  $N(0, \sigma_\epsilon^2)$ . Seasonality is accomplished similarly by learning adjustment parameters for each period of interest (such as afternoon vs. morning or weekend vs weekday) and accounting using the adjust signal for the final experimental analysis.

The key question is how many extra samples we need to account for autocorrelation effects? The degree of information loss is a function of the magnitude of the autocorrelation. A simple heuristic is:

$$n_{\text{eff}} = n\left(1 + \frac{1}{1 - \phi}\right) \quad (7.15)$$

where  $n$  is the original IID sample size.

A simple justification for this heuristic is we are accounting for the influence in subsequent samples of the current sample and under the AR(1) model, this is simply the geometric sequence  $\phi + \phi^2 + \phi^3 = \frac{1}{1-\phi}$ ,  $\phi < 1$ . For  $\phi$  close to 1, the value of the measure would change extremely slowly, requiring an extraordinary number of samples to estimate the true mean and variance.

In practice,  $0.1 < \phi < 0.5$  or between  $1.1n$  and  $2n$  samples. Of course, as mentioned above, auto-correlation can also be resolved by selecting a sampling frequency and strategy that effectively renders sequential samples independent.

Finally, although it was not done for the work in this dissertation, the AR(1) model can be included as a dependency between the input samples to the t-distribution in the primary

trial model using a common  $\phi$  which is given either a uniform or empirical prior.

## 7.4 Confounding

If users report a specific frequency and magnitude of unmeasured or unmeasurable confounders, the system can introduce random alternations to try to balance confounders over multiple phases of the trial. If confounders are deemed to be rare, the mechanisms for adapting an ongoing trial to unexpected events (such as restarting a treatment phase after a delay) may be sufficient.

The acquisition of this model can be driven by user assessment, as well as analysis of baseline data for the outcome across many different experiments and treatments. There appear to be two kinds of confounding influences in the data observed to date:

- **Impulse Confounders.** In Psoriasis patients will experience short term flares if they consume a trigger food. If no more triggers are experienced, the magnitude of the symptoms will revert to the baseline in a consistent pattern. Impulses can be ignored in the final analysis, requiring increased sample size if common, or modeled and adjusted for just like seasonal effects, if confounding factors are clear and can be annotated by users.
- **Frequent Confounders.** By contrast, frequent confounders are like the effect of variations in sleep on fatigue. They happen on a regular enough basis we can't ignore them. If we cannot measure and adjust for the confounding factor as discussed in Section 10.1.1, then the trial requires randomization to ensure that the influence of the confounder is balanced across the baseline and treatment arm.
- **Continuous Confounders.** Continuous confounders are small, always-on effects and can be treated like noise parameters. They have the effect of increasing the variance of the underlying measure. Unless they are clearly identified and can be

controlled for, they will naturally increase the sample size of a trial through their impact on variance.

The above discussion refers to measurable confounders. If large lifestyle changes take place, such as leaving on vacation, a business trip, an extended illness, etc. Then the trial needs to be paused or adjusted as described in Section 8.4.

### 7.4.1 Learning from Special Causes

Confounders are not always bad. Confounders can also be positive indications of a “special cause” influence on the measurable output and suggest new hypotheses to test.

For example, one user of MyIBD discovered while recording data that her worst abdominal bloating always happened on days she was staying up late writing term papers. She thought that it might be the Red Bull she was drinking to stay awake and tried consciously to see if Red Bull did it on another day. After replicating that effect, she removed sugary soda from her diet.

## 7.5 Recommending Experiments

After a number of trials have been run, we have developed an empirical assessment of the probability of an experiment being successful  $P(s_e)$  and the distribution trial’s effect size  $(t|e, s = \mathbf{true})$  if the trial succeeded.

$$P(s_e) = \text{Beta}(\alpha_e, \beta_e) \tag{7.16}$$

$$U(t|e, s = \mathbf{true}) = N(\mu_e, \sigma_e^2) \tag{7.17}$$

The expected utility of an experiment is then:

$$E[U] = P(s_e) * U(t|e, s = \text{true}) \quad (7.18)$$

A rational decision for choosing the next trial to run is maximization of expected utility, or sorted by  $E[U]$ . In the simple model of probability and utility presented here, the expected utility of a trial is simply the quantity:

$$E[U] = \frac{\alpha\mu}{\alpha + \beta} \quad (7.19)$$

Additional refinements can improve this model to account for the relative cost of the treatment  $U(t) = N(\mu, \sigma^2) - C_t$ , the probability of long-term use of the treatment instead of only trial outcomes, and the use of predictive features to personalize the quantity  $P(s_e)$ .



# Chapter 8

## Prototype Design

I developed two web-based platforms, Personal Experiments [ECL11] and MyIBD [ELC12], to evaluate the Aggregated Self-Experiments framework. This chapter expands on the short introduction of Chapter 5 by describing the implemented data model and specific choices for implementing the user and data workflows introduced by Chapter 6. The platform was developed over 3 major iterative cycles, driven by user and collaborator feedback after each prototype was produced.

Personal Experiments hosted two focused user studies to test how users engaged with the concepts and the machinery of self-experimentation in addition to an open observational study of platform use once presented to the public. Results from these studies are discussed throughout Chapter 9.

The design of the platform was heavily informed by an ongoing engagement with the n-of-1 methodology team of the Collaborative Chronic Care Network (C3N) project based at the Cincinnati Children's Hospital and Medical Center. The original code base used for Personal Experiments was expanded under contract to create the second site, MyIBD [ELC12], enabling patient, researcher and clinician triples in the ImproveCareNow [WM11] network to engage in formal and ad-hoc n-of-1 experimentation. The differences between the two deployments is discussed in Section 8.7 and learnings in Section 9.4.3.

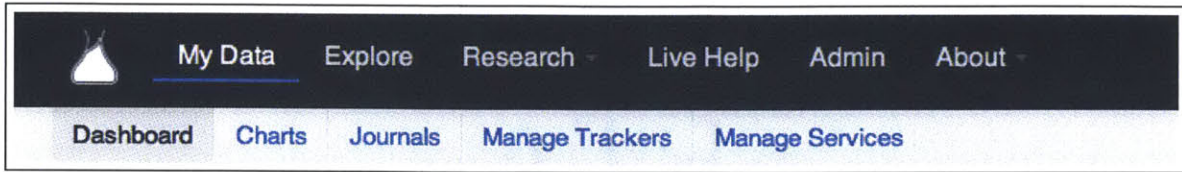


Figure 8-1: Top Navigation Header

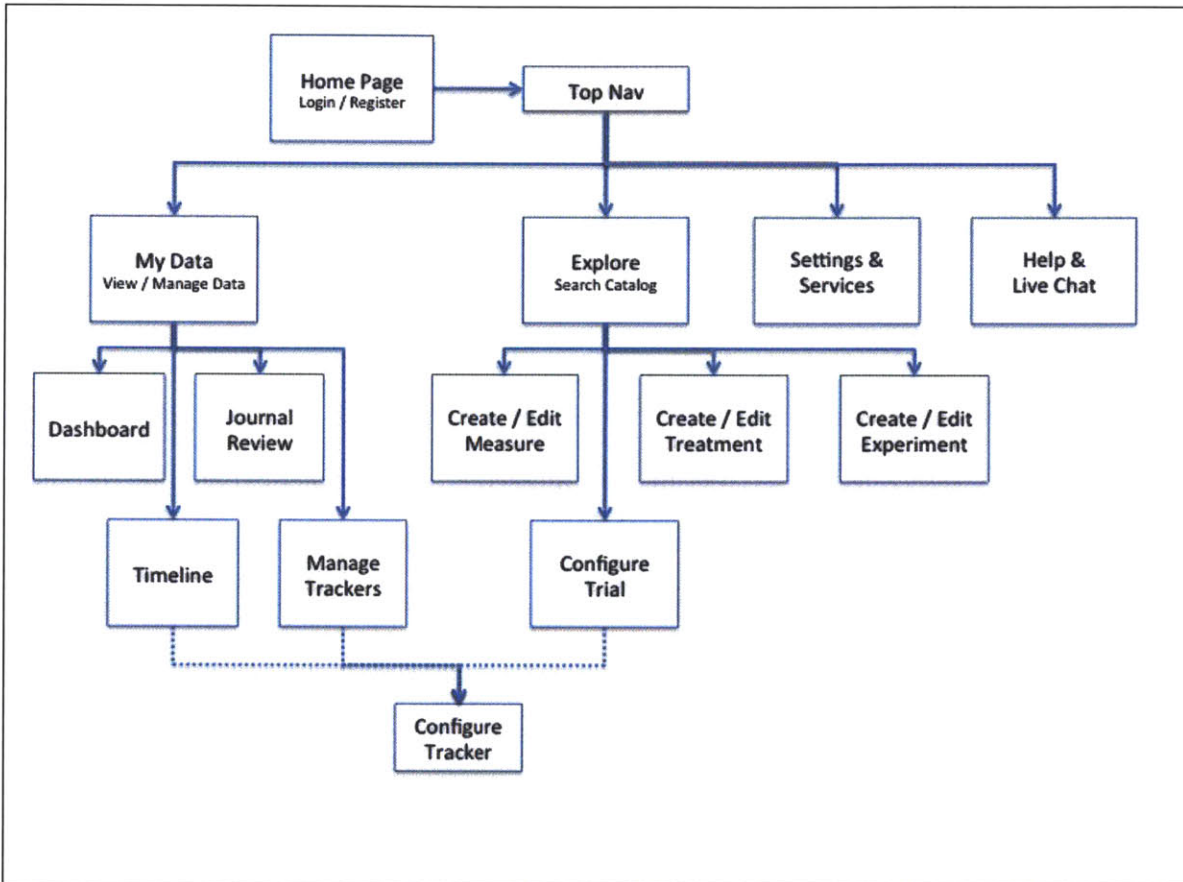


Figure 8-2: Personal Experiments Site Map

## 8.1 Site Navigation

The site consists of four top level navigation states, each with sub-navigation elements as illustrated by the top-nav photo in Figure 8-1 and the site map in Figure 8-2. The Explore interface allows users to search, annotate, create, and edit the catalog of treatments, measures and experiments. The My Data tab provides visualization and interactive features for recording and reviewing the data generated by the system.

The Explore interface (Figure 8-3) helps users find or create the elements described in

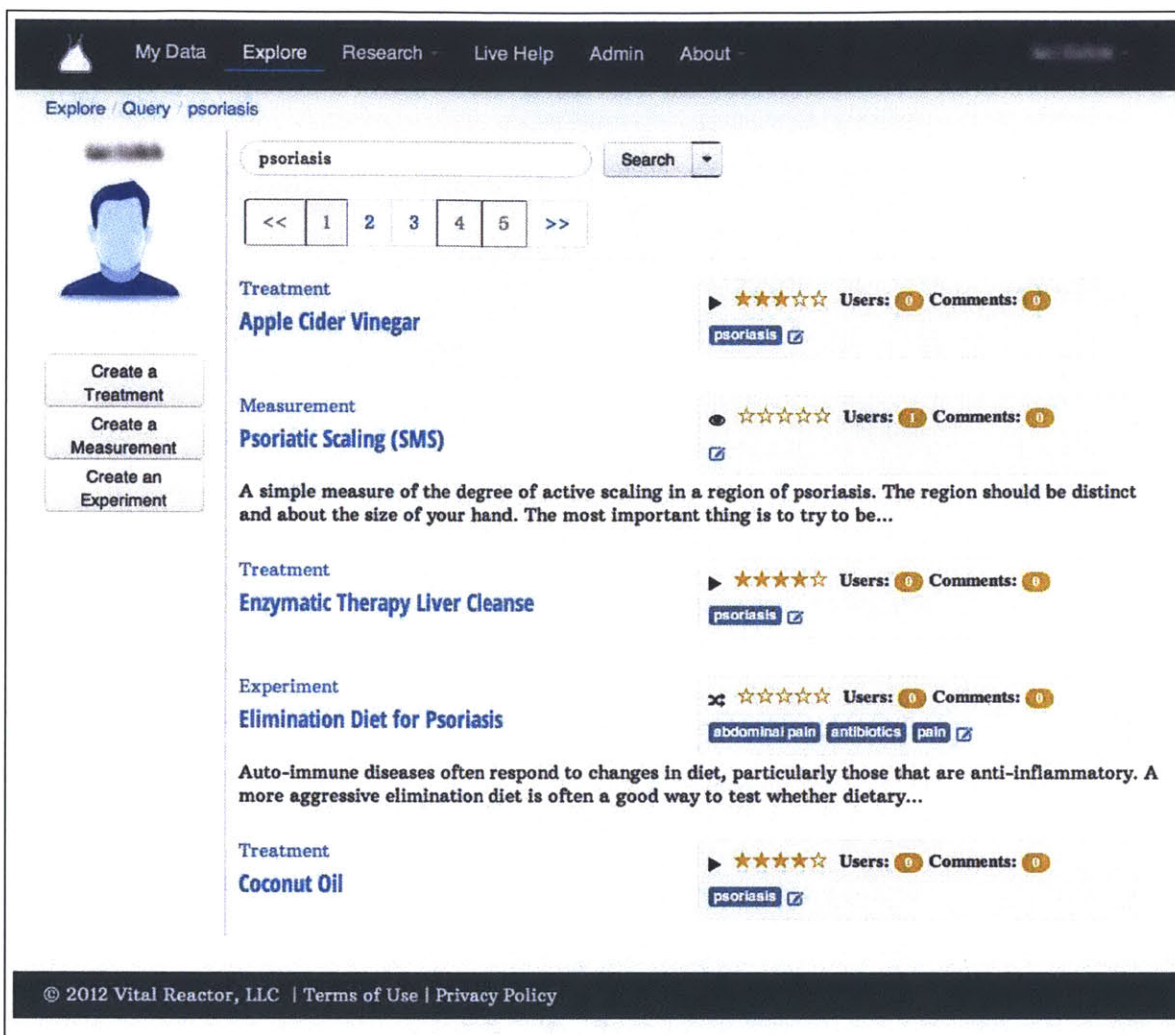


Figure 8-3: Search Interface

Chapter 6. The default user landing page contains the Dashboard interface (Figure 8-4), which shows the current status of trials and all the outstanding actions needed for tracking purposes. Historical data, as well as any experimental context is rendered on the Timeline view under the Charts tab (Figure 8-5).

The remainder of this chapter describes the platform data model and summarizes how these elements facilitate the consumer and producer workflows described in Chapter 6. The final sections describe the MyIBD version of the platform and the technical architecture used by the implementations.

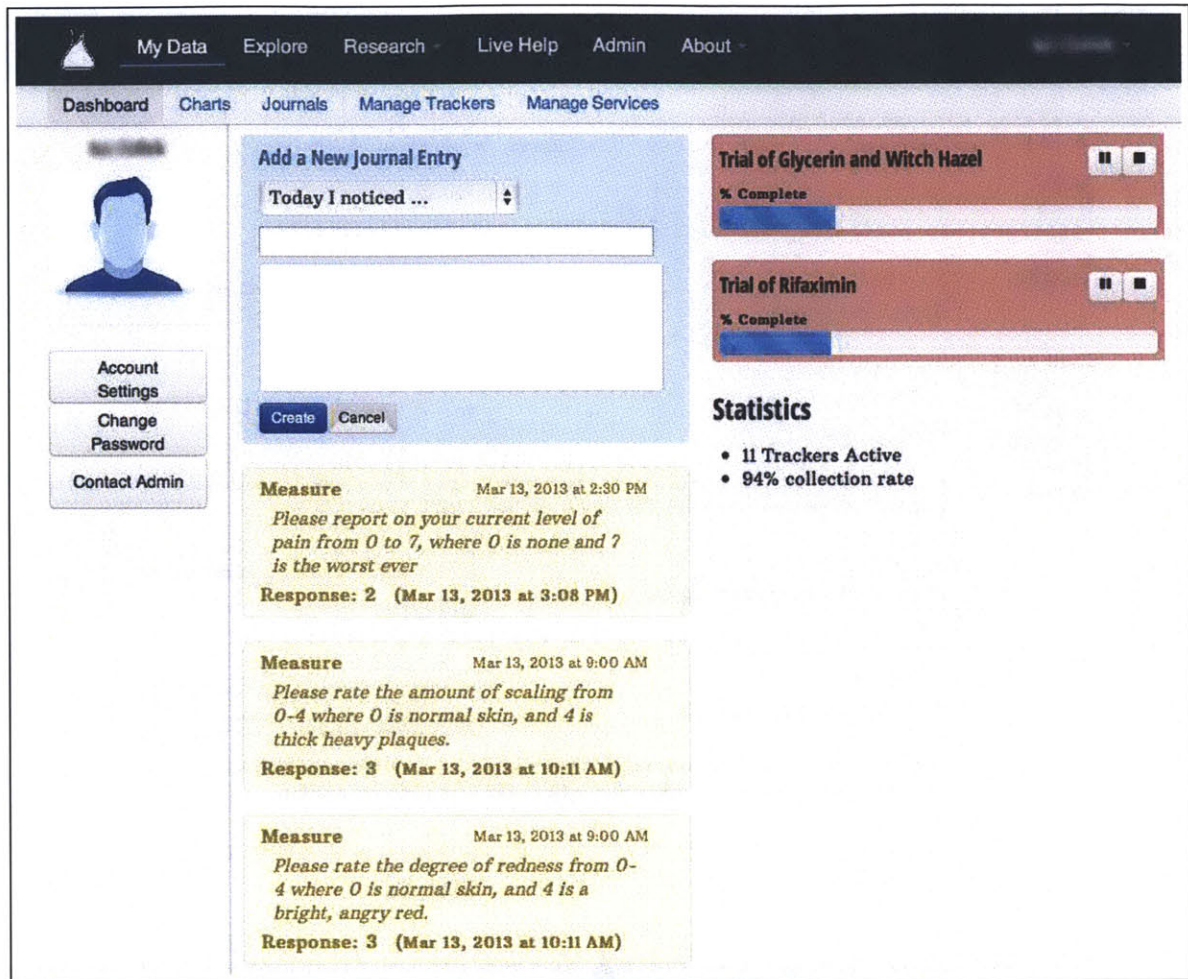


Figure 8-4: Dashboard Page

## 8.2 Data Model

The building blocks of the framework were introduced in Chapter 5; Figure 8-6 diagrams all the core elements of the platform's data schema. A Treatment stores human-consumable treatment descriptions and parameters such as onset and carryover times. A Variable is a text string representing a measurable quantity that can be the outcome measure for an experiment. The method for acquiring samples of that variable from users is represented by a Measure. Experiments link a Treatment and a primary Variable to form a hypothesis that a treatment will change a measure of the variable by a clinically significant amount. Trials and Trackers instantiate Experiments and Measures with parameters such as the trial start date for specific users.

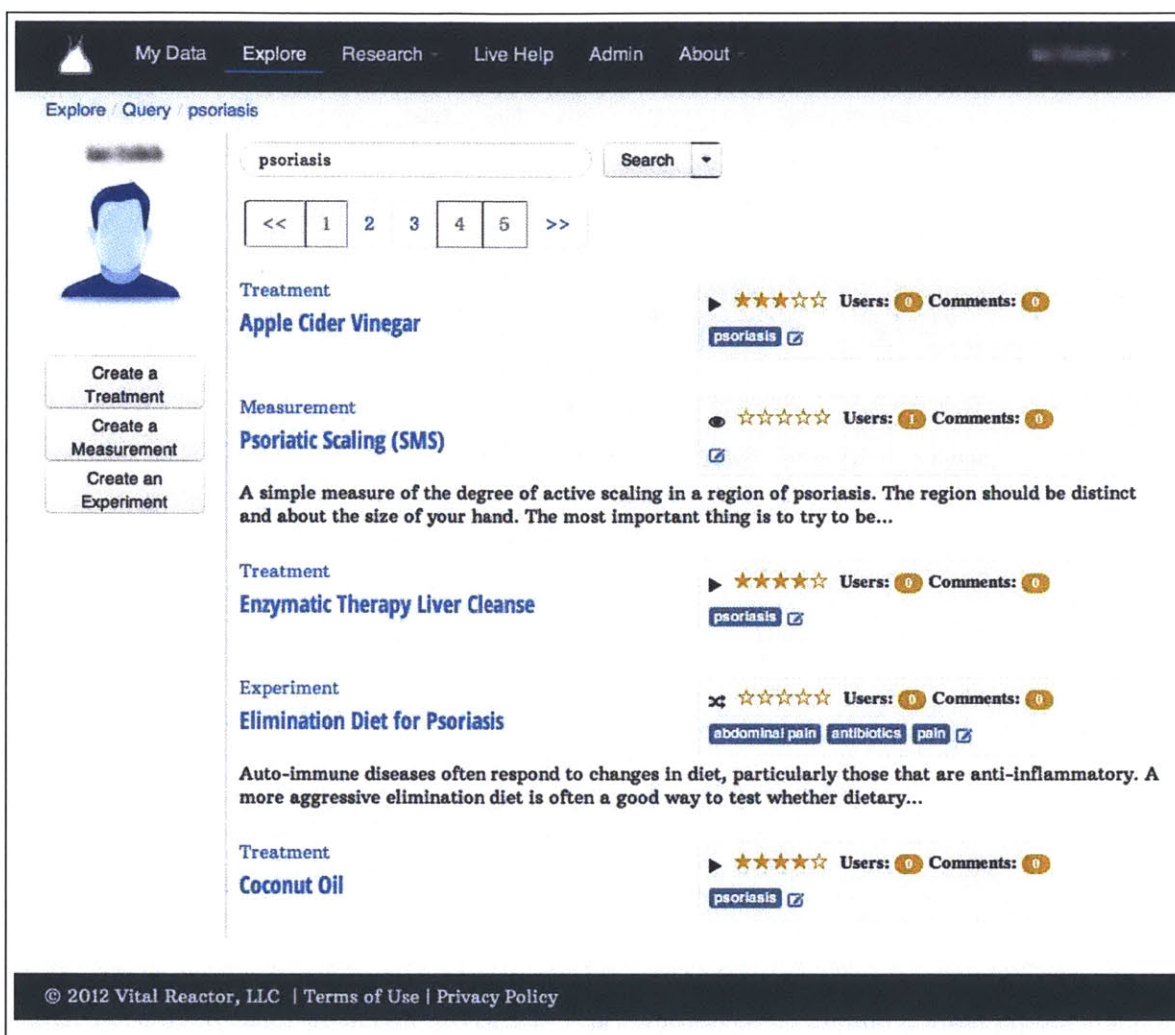


Figure 8-5: Chart Review Page

## 8.2.1 Support for Data Acquisition

The concept of a Variable was introduced late in the iterative design process to dis-intermediate Experiments and Measure and reduce the proliferation of new experiments based on different means by which a single quantity can be measured. The choice of a free text string as the canonical label for a measurable quantity was pragmatic, motivated by the following principles:

1. Medical taxonomies are hard for professionals to use
2. Different communities often use different names for the same quantity
3. The key requirement was that the group agrees on a label

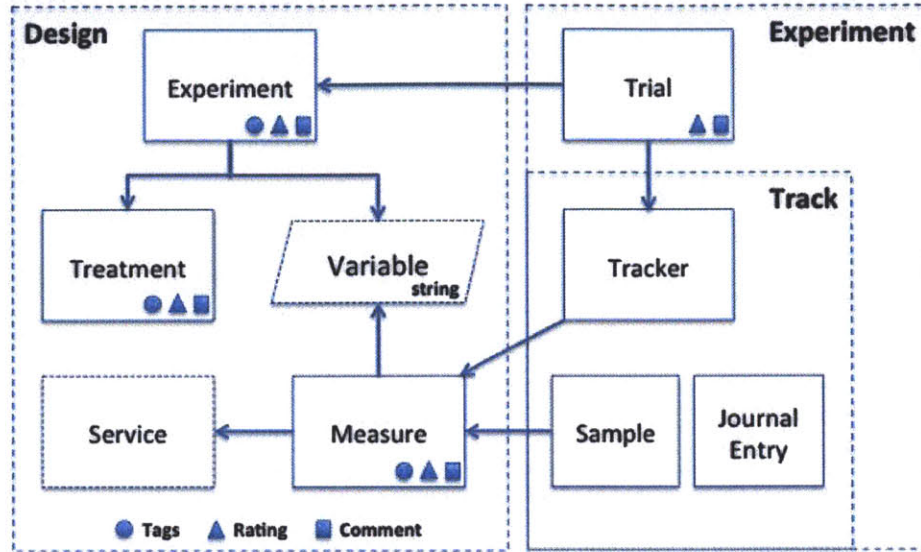


Figure 8-6: High Level Data Schema

4. Auto-completion can reduce duplication in practice
5. It is easy to rename and repurpose labels represented as simple strings


## Measures

A Measure element contains the specification of a specific means of measuring a Variable. It captures syntax, value constraints, guidance to users, prompts, and the type of service (SMS, survey, device, etc) used to acquire samples of it's variable's value (see browser picture in Figure 8-7). A Measure supports the set of fields described by Figure 8-8.

When there are multiple measures defined for a particular variable, it is because there are multiple ways that item can be measured. For example, "hours slept" is a variable that can be measured by a manual estimate over SMS, a Fitbit device, an UP device, or a Zeo sleep meter. There is a concept of a Service dispatches code appropriate to acquiring data for that specific measure. The Fitbit service, for example, fetches an summary record from the server capturing all activities recorded up to a point in the day. Each Fitbit Measure generates time series observations by extracting a single value from that complex record.

Services implemented for the thesis include: Fitbit, Jawbone UP, Rescuetime, Withings,

Explore / View / Measurements / Psoriatic Scaling



Create a Treatment

Create a Measurement

Create an Experiment

## Psoriatic Scaling (SMS)

Untrack
Edit

Details
Discussion

**Description**

A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. The most important thing is to try to be consistent in what the psoriasis looks like from measure to measure.

**Service Type**  
SMS or Email Messaging

**Prompt Text**  
Please rate the amount of scaling from 0-4 where 0 is normal skin, and 4 is thick heavy plaques.

☆☆☆☆☆ Users: 1 Comments: 0

**External References**

[New Reference](#)

**Related**

- ✕ Glycerin and Witch Hazel for Psoriasis
- ✕ Elimination Diet for Psoriasis

Figure 8-7: Measure View

Field	Description
<b>Variable</b>	The variable this measure quantifies.
<b>Service</b>	The service used to acquire the data item.
<b>Description</b>	A free text description of the measurement, often containing instructions or guidance for how to report subjective numbers.
<b>Data Type</b>	The primitive type of data: integer, decimal, categorical, ordinal, or free text
<b>Data Schema</b>	If using categorical or ordinal types, the mapping between values and plottable values
<b>Prompt</b>	(optional) If sending prompts via SMS or another channel, this line contains the actual prompt to send
<b>Reminder</b>	(optional) Most measures support reminders to enter data, this contains a short message to be sent via SMS or e-mail
<b>Domain Min/Max</b>	The minimum and maximum numerical values allowed. This is used both for error checking while parsing user input and for determining the y-axis when plotting.

Figure 8-8: Measure Fields

and Zeo.<sup>1</sup>

There are several open design questions about Measures that remain unresolved:

- Can we estimate a variable value using a combination of trackers? (e.g. to fill in missing items)
- What if we need more than one way to measure a variable using the same service? (e.g. different prompts?)
- For later data analysis, how do we track whether changes were made to prompts that may confound the data?

## Trackers

A user creates a Tracker to support capturing a time series of observations of the Variable according to the template of the Measure object. Trackers are created when the user initiates a Trial, or enables them explicitly on the tracker management page (Figure 8-9). The tracker stores information about the current status of tracking and when to schedule prompts as shown in Figure 8-10.

Service-based trackers will access the service and pull data once a day and only requires service configuration (Figure 8-12). Some services support a push mechanism whereby data is sent from the service to Personal Experiments, allowing data to be tracked in real time. Services using prompts, whether by SMS, e-mail or some other mechanism require configuration of a schedule that suits the user, shown by Figure 8-11. The current implementation is focused on prompts that are sent daily or weekly.

For trackers that have prompts or other scheduled activities the system will call an internal method to spawn a Task appropriate for the specific type of Measure the Tracker supports at the designated wall time in the user's timezone. A Task tracks the state of a specific data collection event. For example, a Measure may support a policy of reminding a user if they haven't responded to a prompt after a certain amount of time. The Task tracks

---

<sup>1</sup>The company behind the Zeo service became defunct during the development period.



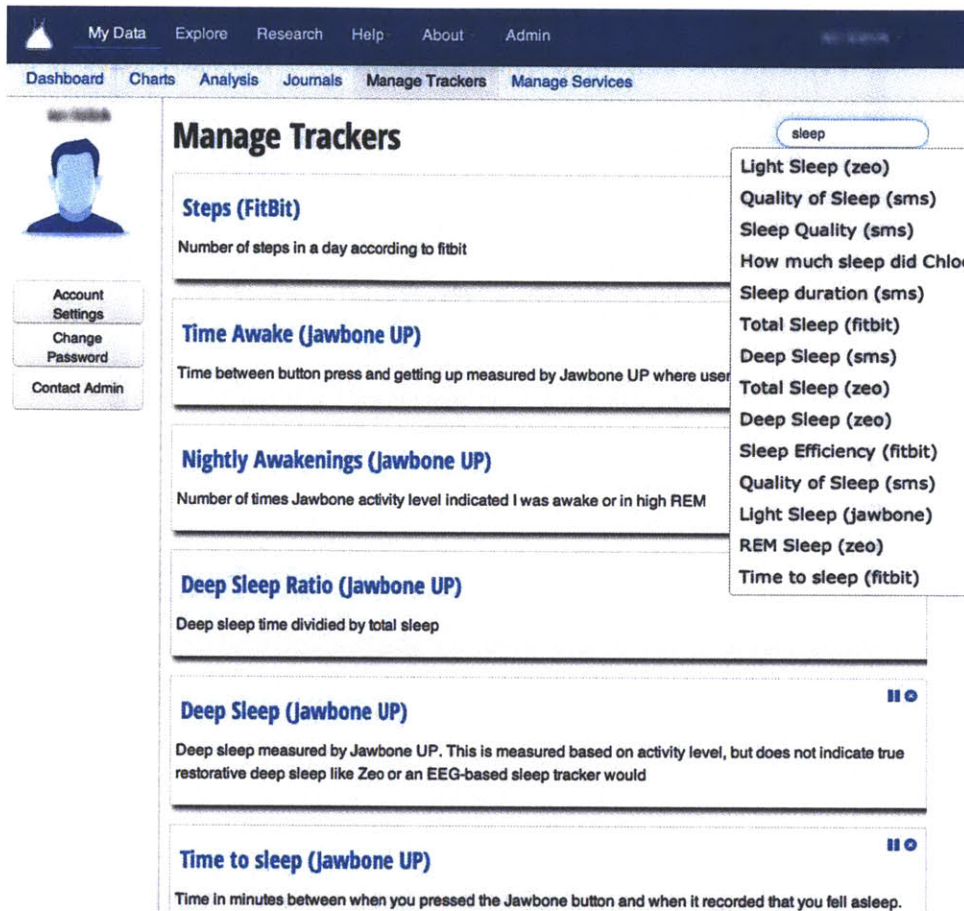


Figure 8-9: Tracker Management

Field	Description
Measure	The Measure this tracker operationalizes.
User	The user this tracker supports.
Active?	A predicate determining if the tracker is enabled or disabled
Schedule	(optional) A specification of one or more, possibly repeating, calendar times for prompting the user for information
Policy	(optional) Several services have a variety of policies that can be employed in collecting data.

Figure 8-10: Tracker Fields

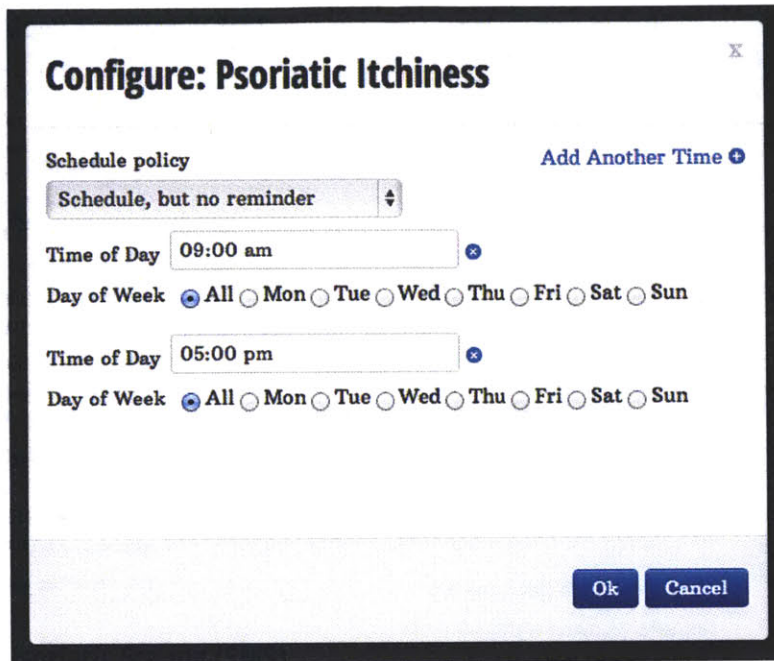


Figure 8-11: Configuring a Tracker

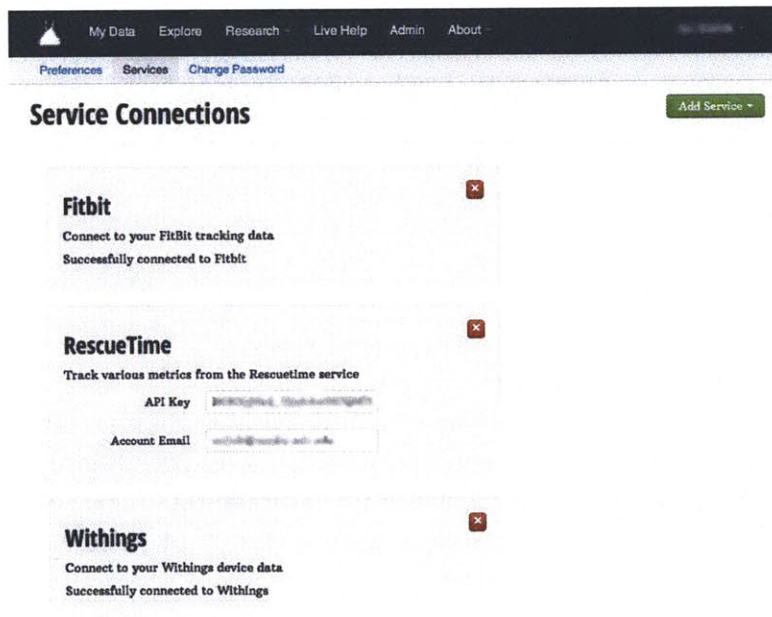


Figure 8-12: Service Configuration

the variables necessary to implement this policy.

Tasks are retained by the system long-term to record the history of events such as time to respond, collection adherence, etc. The user Dashboard (Figure 8-4) uses the database's Task history to highlight the outstanding, completed, or failed tasks for the current user.

### **Tracking via SMS**

The most common form of tracker in Personal Experiment is also the simplest for the user, a text prompt delivered via cellular simple text messaging (SMS or text messaging). The use of SMS is seamless for many users, facilitating remarkably high adherence rates (see results in Section 9.4.3). The SMS Task sends the Measure's prompt to the cellular number entered by the user. An API handler for the user's response SMS looks up the Tracker based on the sending and receiving numbers so it can dispatch the appropriate parser for the measure data type.

The SMS functionality is implemented via the Twilio SMS gateway service that talks directly to the cellular network and exposes a REST API to the Personal Experiments site. One valuable feature provided by Twilio is the ability to dynamically allocate a concrete, persistent cellular number from a given area code to use as the sender of a message. Each tracker for a user is associated with a unique number. This results in a single number used for all prompts from a given Tracker (Figure 8-13), allowing the system to route incoming responses to a parser for the associated Measure. Numbers can be reused across users so the total amount of phone numbers reserved is the maximum number of trackers any user signs up for. Typical users have between 2 and 4 SMS-based trackers.

Another feature offered to users is the ability to annotate data points provided over SMS (e.g. "Hours slept last night?" Response: "8 but woke up 5 times"). The annotation is all the text that flows the response value.) This facility is valuable for capturing additional information that may cause a reviewer to reconsider the system's statistical conclusion because of unexpected factors reported by the user. Annotations also facilitate user recall

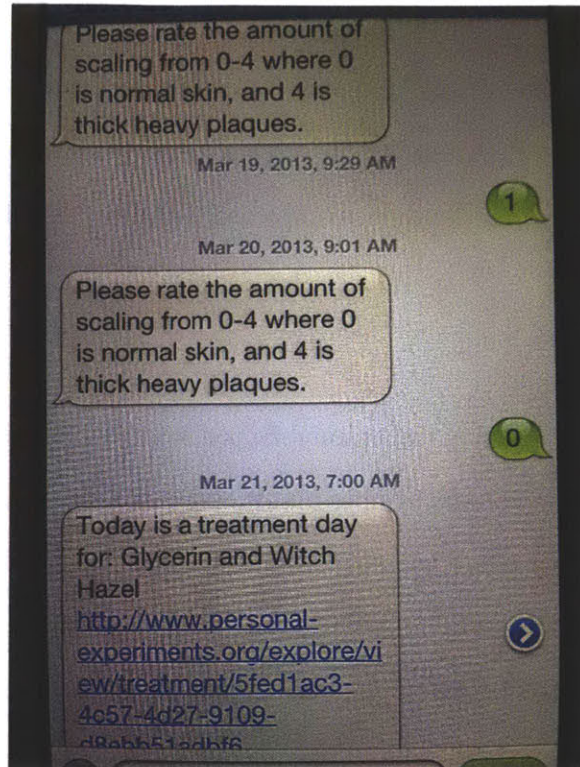


Figure 8-13: Photo of an SMS “Thread” on an iPhone 5

when reviewing history.

Any SMS number used by the system can be used to create longer out-of-band annotations as Journal entries. Journal entries result in annotations on the Timeline view’s x-axis (Figure 8-23) to help interpret or track events for which there are no appropriate trackers. Users simply preface their message with the hash symbol “#”.

### Tracking via 3rd Party Services

As mentioned above, trackers can be introduced to capture any time-series available from a 3rd party service. Services in production at the time of this writing include:

- Withings. A company that provides a wireless scale and an electronic blood pressure cuff that synchronize with their service.
- Rescuetime. A service that tracks second-to-second application focus on a personal computer. With user assistance, it classifies applications and documents into ac-

tivities, categories, and projects. Every use has a productivity score. Time series representing time spent on entertainment, productive tasks, or just total time active at the computer are easy to access via an API.

- **Fitbit.** A pocket accelerometer that measures activity throughout the day and wirelessly uploads it to a service. Some devices produced by the company also estimate total sleep and deep sleep.
- **Jawbone UP.** Another accelerometer-based product worn on the wrist that tracks activity, sleep, and through an associated mobile application diet and mood.
- **Zeo.** Defunct at the time of this writing, Zeo produced a consumer sleep tracking devices based on an EEG detecting headband.

The system registers for real-time push notifications from services like Fitbit and Withings when available or downloads the day's data in the early morning if not. The user can manually refresh to force a fetch of the latest data by clicking a refresh button in the timeline view (Figure 8-23).

## **Samples**

When a value is successfully parsed by a Tracker or Task, it is recorded as a Sample. A Sample is a relational object associating an arbitrarily complex value and metadata with a timestamp, a User, and a Measure. The timestamp represents the reference time for that value. The Sample data usually contains a primary numeric value<sup>2</sup> and the original, possibly more complex value. Sample metadata stored includes how it was collected and the time of day the value was actually recorded.

The benefit of SMS is that the timestamp for when the data was collected is identical to the timestamp associated with the value. However, users can update values for previously completed tasks and provide values for tasks from earlier in the day. Under this scenario

---

<sup>2</sup>Except for free text trackers which aren't implemented in Personal Experiments as of this writing.

the time the data was recorded is different than the time the data refers to. User-visible plots are always generated from the time of reference, but time of recording is available for future analysis.

### **Sample De-duplication and Canonicalization**

The varying semantics of data provided by 3rd party services creates challenges for storing and retrieving time series. The simplest form of data is event based. SMS-based trackers provide event based data at very specific, concrete timestamps that can be used directly in database queries. Others, such as the Fitbit data, summarize accumulated real-time information at the granularity of a calendar day which is not well-represented by databases when the user's time zone must be taken into account. If the user changes timezones, an end-of-day timestamp can end up generating charts in the wrong day.

Service data may be downloaded multiple times or pushed by the service as well as downloaded, it is important to have an explicit de-duplication policy that is enforced at write time. I encountered the need for three such policies.

- **Event series.** Ensure time series consists of unique timestamp and values. If new value for the same timestamp (e.g. a correction) then update the event at that timestamp with the new value.
- **Calendar day summary.** Choose 1 second before midnight UTC as canonical timestamp for storage of a calendar day. One value per timestamp. Always take the most recently provided service value.
- **Timestamp summary.** Summaries are provided for the day, but different intra-day reports have different timestamps. Convert any received timestamps from the user's timezone calendar day to 1 second before midnight UTC for the same calendar day. Maintain unique timestamps. Always take most recently provided service value.

When data that has been converted into canonical form is returned from the database and used to populate charts, it must be correctly translated into an end of day timestamp in

Explore / View / Treatments / Glycerin and Witch Hazel



new 12/18/18

## Glycerin and Witch Hazel

Details Discussion

New Experiment

Edit

Clone

★★★★☆ Users: 1 Comments: 0

psoriasis

**Description**

A topical treatment some patients find help with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied.

Impact on scaling in psoriasis can be seen in responding patients within two weeks, often sooner for other symptoms such as redness and scaling.

**Side Effects**

**Behavior**

days to take effect: 14

days to wash out of your system: 7

**External References**

- Patient Hypothesis  ✕
- Patient Recipe and Places to Buy  ✕
- 2003 Research on Glycerin for Skin  ✕

New Reference

**Related**

✕ Glycerin and Witch Hazel for Psoriasis

Create a Treatment

Create a Measurement

Create an Experiment

Figure 8-14: Treatment View

Field	Description
<b>Title</b>	A short name for the treatment
<b>Description</b>	What is the treatment? Provide typical dosages, how to take, and what it might be good for.
<b>Side Effects</b>	Are there any side effects to be aware of with this treatment?
<b>Reminders</b>	Text to use to remind a user to take the treatment
<b>Units</b>	Hours, days or weeks for onset and offset period?
<b>Onset</b>	How long from the start of the treatment until you would typically see an effect?
<b>Washout</b>	How long after you stop treatment before you return to "normal"?

Figure 8-15: Treatment Fields

the viewing user's timezone.

## 8.2.2 Treatments

Treatment elements record the documentation necessary to use a treatment as well as the onset and washout times (see Figures 8-14 and 8-15).

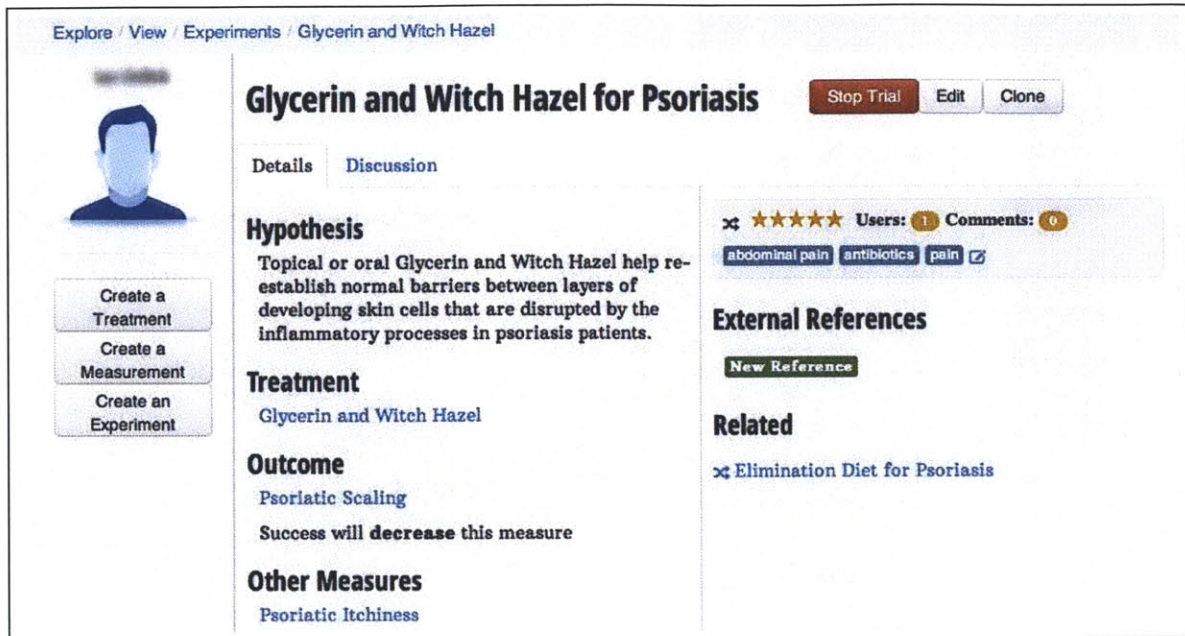


Figure 8-16: Experiment View

Field	Description
<b>Title</b>	A short name for the experiment, often references treatment and measurement
<b>Hypothesis</b>	Why does the creator think the treatment will improve the treatment?
<b>Protocol</b>	Additional information about the experiment. A good place for notes or things to expect that aren't covered in the treatment description.
<b>Treatment</b>	Select an existing treatment
<b>Outcome</b>	Choose from a list of variables for which measures exist
<b>Direction</b>	Are we trying to increase or decrease the measure.
<b>Covariates</b>	A list of other variables to measure for context or future multivariate adjustment.

Figure 8-17: Experiment Fields

### 8.2.3 Experiments and Trials

Experiments combine a Treatment, a Variable, and a human readable protocol to follow to evaluate the impact of the treatment on the primary outcome. A catalog view of an experiment is found in Figure 8-16 and the available fields of the experiment are detailed in Figure 8-17.

Not all measures for a variable are required to be of the same data type, but the cur-



Field	Description
<b>Units</b>	Hours, days or weeks?
<b>Baseline</b>	The number of units for the initial baseline period. May be longer than subsequent periods due to accommodation of practice effects, etc.
<b>Period</b>	How many days per phase to estimate the mean outcome measure for comparison?
<b>Randomize?</b>	Randomize ordering of treatment/baseline within a single cycle.
<b>Cycles</b>	Number of treatment/baseline pairs.
<b>Onset</b>	Override default treatment onset period
<b>Washout</b>	Override default treatment washout period

Figure 8-18: Advanced Experiment Fields

rent Experiment element does not work for non-numeric outcomes. A system designed to integrate with a medical records system may need to add more constraints and an explicit medical terminology. Given user response to the system (Section 9.2.3), it is clear that the complexity barriers created in asking users to use and extend a medical taxonomy would outweigh any benefit of better standardization. Instead, I advocate relying on a community of users to annotate measures with ontology labels as needed for integration with foreign systems.

An advanced editing mode is available for users to explicitly set all the parameters of an  $A(BA)^+$  style design, otherwise they are chosen behind the scenes according to the sample size prediction and adaptation methods laid out in Chapter 7. The extra fields available in the advanced mode are shown in Figure 8-18.

Operationalizing an experiment involves creating a trial of that experiment. The trial creation process involves choosing specific measures from a list of possible measures, deciding on reminders, and configuring reminders and trackers according to any instructions in the trial protocol. The graphical creation view is shown in Figure 8-19 and the details of the fields in Figure 8-20.<sup>3</sup>

---

<sup>3</sup>One potential drawback of the current tracker model is that trial configuration relies on the user to configure a measure schedule according to the trial protocol during trial creation. It is likely that the current method of creating trials is confusing for some users, and an incremental engagement strategy of starting a trial immediately with reasonable defaults and then answering questions to slowly modify the trial to their preferences would improve engagement and adherence.

## Start a trial of: Glycerin and Witch Hazel for Psoriasis

Advanced View

### Trial Instructions

Apply treatment during the treatment periods only, use regular care in the baseline periods. No other special instructions needed.

### Treatment Description

A topical treatment that some patients find help with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied.

Impact on scaling, redness and inflamed area in psoriasis can be seen in responding patients within two-three weeks, often sooner for other symptoms such as redness and scaling. Itching often subsides within days.

### Potential Side Effects Glycerin and Witch Hazel

Witch hazel with alcohol can sting on application. Application of this mixture to open sores (e.g. after scratching) can hurt a little. No other harms or side effects are known.

### Configure your trial

When do you want to start?

13 March 2013

Do you want treatment reminders?

Yes  No

### Select your measurements

#### Psoriatic Scaling

via sms

#### Psoriatic Itchiness

via sms

#### Psoriatic Redness

via sms

### Configure trackers

#### Reminders for 'Glycerin and Witch Hazel for Psoriasis'

Please don't forget to apply your glycerin/witch hazel mix

Reminders via SMS Daily at 10:00 am

#### Psoriatic Scaling (SMS)

A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. The most important thing is to try to be consistent in what the psoriasis looks like from measure to measure.

Create

Cancel

Figure 8-19: Create a Trial

Field	Description
Start	The date to start baseline data collection.
Reminders?	Configure treatment reminders during treatment periods?
Trackers	Choose specific measures for the trial outcome and covariates from a list of possible measures (Figure 8-20)
Phases	Internally, the Trial maintains a list of all distinct phases of the trial design.

Figure 8-20: Trial Fields

Field	Description
Effect Size	“What size of an improvement would satisfy you?”
Confidence	“What confidence level do we need to reach to consider the trial a success?”
Existing Data	Yes/No. If existing outcome data is available, automatically use it for the baseline data. (default = Yes)

Figure 8-21: Advanced Trial Fields

Like an experiment, the trial also has an advanced mode that allows interested users to adjust the specific trial design to accommodate different goals identified in Figure 8-21. This provides more customization when the trial design is not pre-determined by time selections in the advanced mode to exchange time and confidence. Median values for all parameters are chosen by default.

The effect size choices are based on the classical conventions for interpretation of Cohen’s  $d = \bar{X}_2 - \bar{X}_1/\sigma$  (refer to [Coh88] or Section 7.2.2) as small, moderate, and large improvements. Future versions of the framework could select these levels based on user feedback about perceived improvement. Regardless, the purpose of the effect size is to ensure that there is a standard notion of effect size, and that a trial will collect enough data to detect an effect at least as big as the users preference.

A target confidence factor establishes the users tolerance for false positives:

1. **More likely than not** Greater than 50% probability that treatment is better than the baseline,
2. **Likely** 80% probability that the treatment is better than the baseline, or

### 3. **Very Likely** 95% probability that treatment is better than baseline.

After the trial is launched, a progress summary widget is added to the user's dashboard (see Section 8.4 below).

Internally, the trial consists of a sequence of time intervals each associated with a baseline or a treatment. The duration of each phase is computed as described in Chapter 7. These phases are used to populate the control charts shown earlier and to track changes in the trial as it evolves. Trial reminders look at the current state of the trial's phases to assess what reminder to send to the user.

## 8.2.4 Miscellaneous

Annotations, as described previously, are simple text comments added to data points. Journal entries consist of a short title and long body and describe some aspect of the user's state not captured by existing measures. Journals and data annotation have received a considerable use by users, particularly in MyIBD where users are documenting their experience to provide context for their tracking data to their clinicians.

The site also provides facilities for adding references, comments, tags, and ratings to any of the three major objects. These mechanisms were introduced to allow individuals to share information and opinions anonymously. They have not seen much use at the time of this writing.

## 8.3 Consumer Tracking Workflow

A typical consumer starts with the Explore search interface (Figure 8-3). They enter any symptoms or diagnoses they have, or treatments they may have heard about. The search keywords identify matching elements, including hierarchical relationships such as returning experiments for treatments that have tags or title terms matching the search query. Users can restrict types by prefixing "show type" to the search query where type can be measure,



Figure 8-22: Search Result Detail

treatment, or instrument. A dropdown next to the search button allows users to browse all objects of a given type.

Each search result consist of type, title, and description information. It also provides a badge that captures the aggregated information about the result (Figure 8-22). The badge has a glyph that indicates the kind of object, followed by a 1-5 star rating which characterizes either the accuracy of a measure, effectiveness of a treatment, or quality of an experiment. Users can hover over the stars to get a tooltip describing what the stars mean, and can click to provide their own rating.

The badge also contains the number of users who are using or have commented on the element<sup>4</sup> and for measures, a “validated” tag identifies measures that have been formally evaluated in the literature (not shown in Figure 8-22. The bottom half of the badge consists of a user-editable list of tags that represents an informal taxonomy of concepts associated with the element. This typically includes information such as diagnoses or general symptoms that aren’t part of the formal data model described above.

### 8.3.1 Tracking

When the user finds a Measure they are interested in, they can click on the “Track” button (shown in Figure ??), or if already tracking, they can click on “Untrack”. The tracker configuration dialog comes up for configurable measures to allow the user to select a prompt schedule and policy for the tracker. New trackers populate a chart review page (Figure 8-

<sup>4</sup>The number for experiments and treatments is the number of past or current trials using either, and for measures the count is the number of currently active trackers.

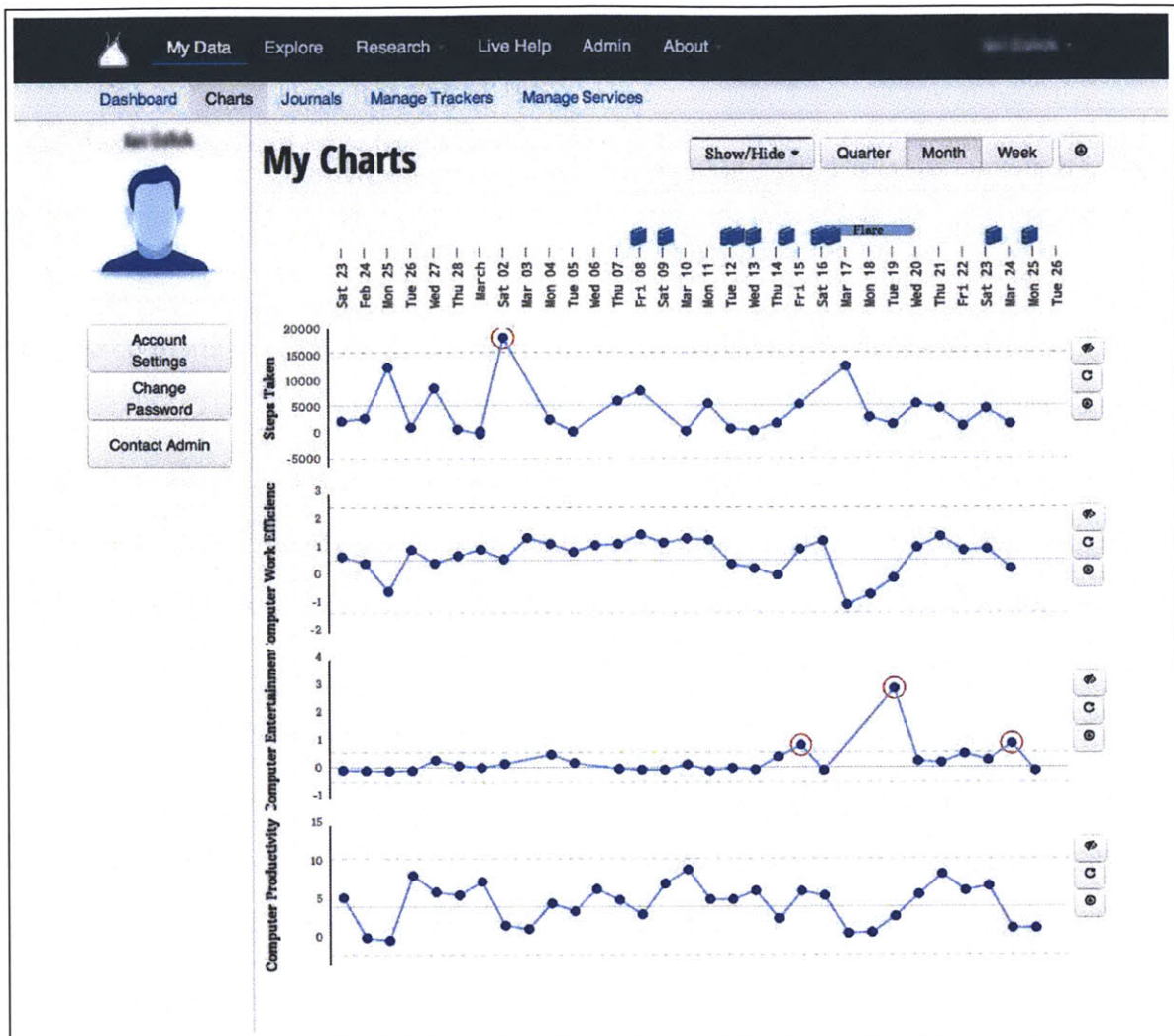


Figure 8-23: My Charts Page

23) with a run chart (see Section 7) that plots the collected data points as they are received.

The top of the chart view consists of a shared context which consists of the date-based x-axis, glyphs for each day's journal entries, and bars associated with interval-based journals (typically used to report medications). Individual data points with free-text annotations are highlighted. Moving the user's mouse over any data element results in additional detail being presented, such as exactly value and annotations for data points.

The chart view allows the user to select different window-sizes for the visible plot domain, either a week, a month, or a quarter (3 months). The entire chart view is horizontally scrollable to facilitate the review of data beyond the 3 month mark.

### 8.3.2 Interpreting Evidence

When sufficient data is available to estimate the mean and variance (see Section ??), the run chart is augmented with a mean measure and upper/lower control limits. The control limits help the user identify points that fall within normal variation. Points that fall outside the control limits are considered “unusual” and tell the user that something new is causing the change from the historical norm.

Run charts are appropriate for measures that represent a measure of the state of a system and is expected to remain relatively stable over long time periods. Measures like weight or fat mass that we aim to steadily change in one direction throughout an experimental intervention require different statistical models, such as a change in slope.

The control limit calculation is provided to facilitate user inference. In the tracking context, significant changes suggest that a special cause was present that pushed the normal measure off its stable range. The user can review their journal entries or appeal to memory to try to identify the likely “special cause” that pushed the measure outside of its normal pattern of behavior. This cause may in fact be evidence of a trigger that worsens the measure, or a practice that improves it.

Shewhart originally recommended a three-sigma control limit to balance the cost of investigating possible special causes in industrial systems vs the cost of missing beneficial causes with smaller effect sizes [She39]. In Personal Experiments, however, the inference is intended solely to support user investigation of hypotheses that might motivate further investigation. The consequence of a false positive are low whereas the consequence of a false negative can be high as it risks losing engagement when clear shifts are present visually, but go undetected by the system. Therefore, less conservative control limits are called for. But what limits should be chosen?

Another constraint on the choice of control limits is the nature of the measures. While some measures consist of real-time count data such as steps or hours of sleep, many measures in Personal Experiments are 0-5 or 0-10 metric scales capturing a subjective assess-



Figure 8-24: Dashboard Trial Widget: Active Trial

ment of a continuous quantity. The quantization of these measures mean that three sigma control limits are often fall at the extremum of the scale making special cause detection in this manner impossible. Control limits for these measure are better selected and coupled with decision rules that capture short runs of data outside the limits, rather than allowing for no detection at all.

The control limit rules implemented are:

1. **Short-scale Integer Data:** 2-sigma control limits adjusted to be just below or above the nearest integer measure. The decision rule is two sequential points outside the control limits. Due to the adjustment, a detection event implies 80% to 95% confidence in the presence of a special cause at or prior to the detection.
2. **Continuous data:** 2.5-sigma limits with single-point detection. A point outside the limits represents a 90% confidence in a special cause.

Over time, especially if beneficial treatments are being tried, the baseline for a given measure may change. The scheduling user interface allows the user to select a new baseline start date to reset the chart limits. Section 10.6.3 discusses the prospects for automatic adjustment of run-chart limits in future versions.

## 8.4 Consumer Experimentation Workflow

When the user wants to try an experiment, they can click to configure a trial as described in Section 8.2.3. Once the trial starts, the dashboard trials widget (Figure 8-24) shows the current percent of the trial time that has been completed and exposes action buttons.



### 8.4.1 Trial Actions

The action buttons allow the user to modify the trial mid-stream due to unanticipated circumstances. For example, the user may become sick which is a special cause that would invalidate the phase of the trial it occurs in. The user can pause the trial at a point in time and un-pause later. Pausing a trial pauses all data collection events and reminders as well. A pause of up to 2 days is treated as missing data.

A pause longer than 2 days can initiate a complex decision process about how to adapt the trial to unexpected circumstances without abandoning the collected data entirely. For example, are we in a baseline or treatment phase? If a treatment phase, was the treatment maintained during the paused period? The uncertainty of the user's state increases with the length of the pause. While accommodation can be made to modify the phase length, the simplest is to restart the current phase.

If a treatment phase, the user is prompted to restart the treatment. The onset period starts immediately regardless of whether the user had maintained the treatment during the pause. If in a baseline period, the user is asked whether they took any treatments that might effect their outcome or if they are unsure. If yes or unsure, the entire baseline period including washout period is repeated. Otherwise, just the minimal sample size needed to re-establish the baseline mean for the current phase is repeated.

Internal to the system, the old phase is marked as incomplete and the period it represents is ignored for analytical purposes. A new phase is created of the appropriate type and the trial completion percentage is adjusted accordingly. Thus the completion measure is non-monotonic and represents the current best forecast for completion.

The user can also choose to abandon a trial. When they click on a dialog to indicate they are sure they want to cancel the current trial, they are asked to respond as to why they are terminating the trial early (as Alice did with the Glycerin trial in Section 5.4.1). The dropout rationale assessed include:

1. Tracking was too hard to maintain

2. The treatment was too hard to maintain
3. I believe the treatment worked, but I don't want to keep doing it
4. I believe the treatment worked, I don't want to stop doing it
5. I believe the treatment did not work, I don't want to spend more time
6. Something came up, may try again later
7. I just don't care about experimenting
8. Something else...

Each one of these options supports interesting future inferences about the user and the component elements of the trial. The first three, for example, tell us about the user's dissatisfaction with the measure or the treatment. The early negative response (#5) implies that the treatment does not have a reliably strong effect or was negatively confounded, so the platform asks whether the user is confident that nothing else might have come up to obscure the treatment's effect before accepting the abandonment. The user can also opt to add another experimental cycle to increase confidence.

If something came up, as in #6, then the system engages in a short dialog to ask whether they want to retry soon and use some of the data they already collected (if in phase 2 or later) or come back another time.

If the user says they don't care about experimentation (#7), that is useful feedback on burden of the platform itself and if something else (#8), it asks for a further comment from the user.

If a user gets a dramatic response and refuses to stop taking the treatment (#4), should that be considered a success or a failure? In a traditional trial this is considered a dropout event, but the reason for the dropout is important and the trial abstraction should not keep us from learning from this particular user's experience. Treatments that the user think worked and don't want to stop either have a large subjective effect size, or some confounding factor such as the placebo response, confirmation bias, or another external factor caused the user to believe that this was the case. The system asks for approval to follow-up with the user

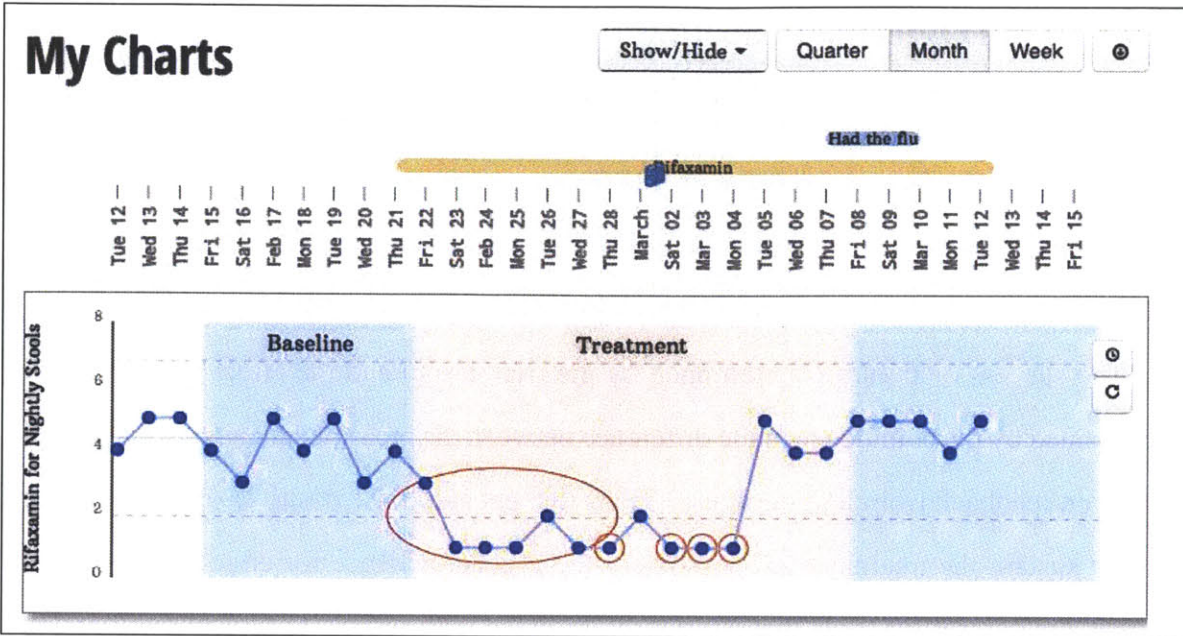


Figure 8-25: Timeline Control Chart

later to see if the effect remains.

### 8.4.2 Visualizing Trial Data

The trial’s control charts are available on the timeline page as shown in Figure 8-25. The control chart plots the trial outcome measure against the baseline and treatment phases determined by the trial design. Additional decision rules and visual annotations are implemented to catch shifts in the mean that do not produce outliers beyond the control limits or for measures where there is a shift in the mean, but the treatment variance is lower than the baseline variance.

I opted not to blind intermediate data; users can choose to review the trial chart state at any time. I also opted to display significant event detection during the trial, but it would be easy to blind either the data and/or interpretation of the data until after the trial was complete. This would include some combination of hiding detection events, hiding the trial review page, pulling the control charts from the My Charts timeline, and pulling the outcome run chart from the timeline.

### 8.4.3 Trial Outcome Reporting

When the trial is completed, the trial status widget is modified to communicate:

1. Was the trial a Success or Failure? ( $P(\text{Effect}_{obs} > \text{Effect}_{targ}) > \text{Confidence}_{targ}$ )
2. Observed effect size in absolute units (“Average value increased/decreased by 1.2”)
3. Confidence that the treatment really works for the user (Low, Medium, High)?

The success or failure is determined by the trial design and the effect size is entirely determined by the confidence in the difference between the measured means. Confidence is based on feedback from long-term use. Trials that get easy false positives, but where users do not sustain the treatment, are less likely to be good evidence for effective treatments. This is the control for trials that are poorly designed with respect to typical confounders, treatments and outcomes with strong placebo effects, or users with a tendency to confirmation bias. The specific computation of these factors was described in Section ??.

The user can choose to archive the trial to de-clutter their dashboard any time after the trial is complete.

### 8.4.4 Successful Treatments and Abandoned Trials

What should we do when a user abandons a trial because they believe the treatment works too well? Assuming the treatment phase was completed and the trial was not explicit, the statistical analysis is switched from the default withdrawal design to the interrupted time series design and re-evaluated as success or failure. This choice may be controversial, yet it does no good to argue against a user’s subjective experience on formal grounds if there is real evidence of effect.

The risk of the re-analysis is that the trial may tend to false positives for the reasons already mentioned, if assessment of trial quality is improved by the success, and more users try it, they waste their time. Controlling for this runaway condition is capturing the long-term success of the treatment, that is if no one sustains the treatment despite successful treatments, then the treatment is de-emphasized. Trial success is an imperfect proxy

measure for true treatment efficacy.

## **8.5 Collaboration Features**

The site provides a variety of collaborative features to facilitate collective intelligence style aggregation and direct peer review and community interaction. As described in Section 6.3, user who fit a producer profile are empowered to create new treatments, experiments, and measures to expand the space of hypotheses that they and consumer can explore. The features in this section support producer-to-producer and produce-to-consumer interactions to facilitate critique and improvement of experiment design and treatment protocols.

### **8.5.1 Discussion**

Each element of the system can be discussed by users through the discussion tab in the Explore view or at the bottom of the trial review page. The current discussion model is a simple linear sequence of a single text comment of arbitrary size. Existing comments can be edited by their creator.

### **8.5.2 Ratings**

As mentioned above, each element visible through the Explore tab provides an interactive 5-star ratings widget. Users provide a subjective assessment of the value of a given element and the mean score is reported there. As empirical evidence is accumulated for or against the effectiveness of a treatment or success of an experiment, this subjective 'prior' is de-emphasized in favor of empirical data.

As personalized predictors are available, the star ratings are customized for each treatment and experiment to indicate the likely value/outcome for them. Presenting these various elements independently, and evaluating their impact on users, is a good topic for future research (see Section ??).

### **8.5.3 Trial Peer Review**

Trials can be designated as public, protected, or private. Public trials are visible in a time ordered list in the explore tab's experiment view (Figure 8-16). Clicking on one of these trials will pull up the trial review page described above. Free-text discussion can take place against this trial. Public trials have unique URLs based on their database UUID, allowing user to share trials with other users.

A protected version of a trial is also planned, meaning that trials are only listed in experiment views when being viewed by designated "producer" accounts or when navigated to directly via a shared URL. This facilitates peer-to-peer sharing and professional review without exposing any personal data to the public internet.

### **8.5.4 Experiment Peer Review**

Experiment peer review is primarily facilitated through the public discussion widget on the experiment-view page (Figure 8-16). However, a future version will add population statistics as another view tab allowing the aggregate population experience to be viewed along with the results of individual trials marked as public or protected (for producer users). For now, I play this role in a "Wizard of Oz" [DJA93] style by manually reviewing data using the system command line and direct engagement with users via the admin e-mail account, live chat interface, and discussion widget.

## **8.6 Aggregation Support**

Limited aggregation support is built into the prototype system, although due to the small number of data points collected at the time of writing, was not fully integrated into the user interface design. As shown in Figure 8-30 below, updates to objects in the database, including all new samples submitted, trigger an internal statistics update to modify the system's prior distributions for measure variance, probability of experiment success, and

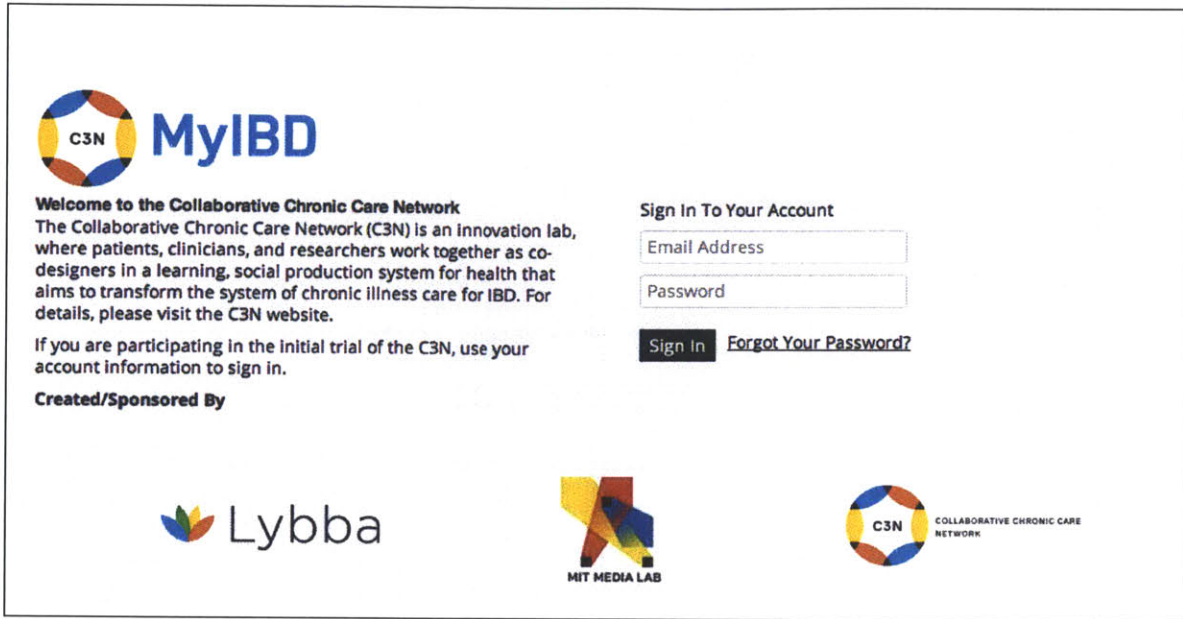


Figure 8-26: MyIBD Login Page

expected effect size.

The default experiment model leverages these parameters to calculate trial phase durations.

## 8.7 MyIBD Deployment

The MyIBD version of the site (Figure 8-26) plays host to an ongoing 50 patient study of active learning in a clinical care setting. Older pediatric patients with inflammatory bowel disease such as Crohn's or Ulcerative Colitis use the platform to track longitudinal Patient Reported Outcomes (PROs). The system in the field uses a learning plan form (Figure 8-27) and the ranged-journal reporting to capture patient goals and define more formal experimental interventions.


Clinicians and all members of the methodology and statistics team have de-identified access to the patient user population to facilitate peer review, shown in Figure 8-28.<sup>5</sup> Clini-

<sup>5</sup>There is a similar administrative page for administrator accounts. This view enables the clinic's administrator to create, review, and edit all users with identifiers. They can use the tools to contact individual users or the entire population by e-mail or SMS, reset passwords. A planned future release adds a clinician role and configuration screen allowing the administrator to assign patients to clinicians who will then be able to

**MyIBD** My Data Explore Research Admin Ian Eslick ▾

Patient List Charts Journals **Learning Plan** Trackers

**Demos Userman**



**Diagnosis**  
Psoriasis, Dysbiosis

**Short Bio**  
An adult demo user; has a fake bio and history, but contains a copy of real-world data collected on this site.

[Send a Test SMS](#)

## My Learning Plan [Edit Plan](#)

**Learning Plan**

<i>Title</i>	A quest to improve energy and mental acuity
<i>Goals</i>	My primary goals are improving my energy levels and mental acuity. Secondly I'd like to reduce the intensity of my psoriasis symptoms.
<i>Background</i>	Ian was diagnosed with psoriasis at the age of 13. Early diagnosis is a risk factor for early onset co-morbidities such as stroke, heart disease, etc. His lipid levels are high-normal, but he had a strong inflammatory response to a recent Cleveland blood draw indicating risk of stroke. At age 27 his psoriasis increased markedly along with the onset of symptoms related to energy and mental acuity. In the recent years he has managed this through a restricted, low-carbohydrate diet with minimally processed foods. He regularly supplements with B-vitamins and natural anti-oxidants.
<i>Potential Interventions</i>	Ian is interested in the GAPS and Paleo diet and the reported success of some people in treating auto-immune diseases.

Figure 8-27: MyIBD Learning Plan Page

cians or reviewers can choose a patient context then view/edit the patient's charts, journals, and trackers using the same views employed by the patients themselves as shown by Figure 8-29. The other pages available are identical to what individual users have available in Personal Experiments.

One of the central design objectives was to enable as much transparency as possible under legal, regulatory and practical restrictions. Users view and use all the same tools that clinicians use, can add their own trackers, adjust their own schedules, and will eventually co-create experiments and run them independently. During clinic visits, both parties look at the same screens to discuss the patient's data and adjust plans. The system retains a history of all the changes made and associates each change with the user that performed them. This data can be exported from the system for offline review.

Interaction with patients to date primarily involves symptom surveillance. Surveillance enables clinician and patient pairs to document, discuss, and hypothesize about variation in

---

see their own patients in their fully-identified form, including photos instead of generic icons.



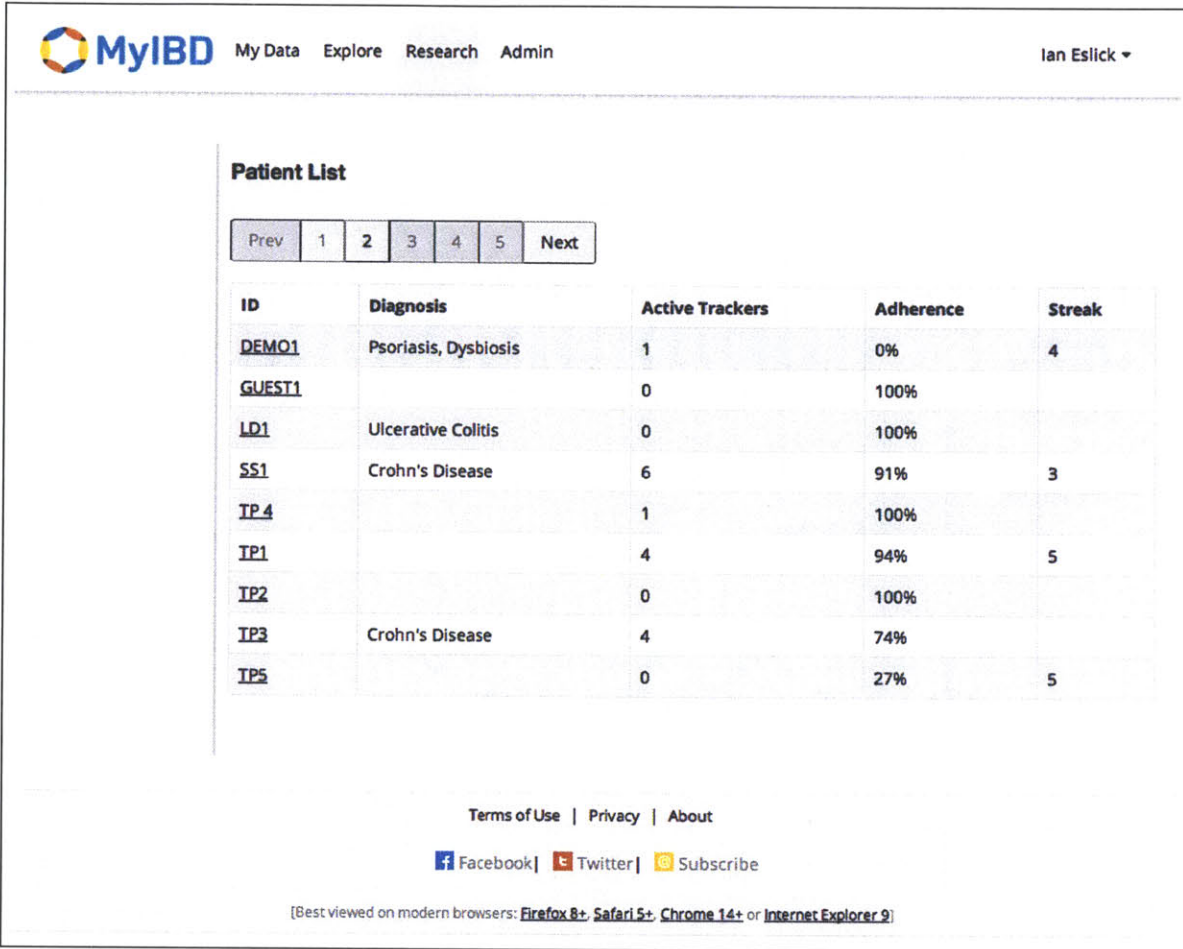


Figure 8-28: MyIBD De-identified Population Browser

symptoms over time. This supports a search for special causes (such as trigger foods). Most IBD patients have regular clinic visits along with tests for primary disease activity, but many IBD patients have symptoms that dramatically impact quality of life yet are uncorrelated to these measures of disease activity [MPS13]. These symptoms can involve phenomenon like nighttime stooling (causing frequent awakening), fecal or urinary urgency (makes it difficult to leave home), bloating and abdominal pain (due to bacterial activity in the small bowel or illiim), etc.

Unfortunately, there is little scientific evidence for or against the alternative or lifestyle interventions parents and patients often use to attempt to manage these symptoms. The clinician is often at loose ends as there is no clinical evidence base to work from. The team hypothesized that tracking and n-of-1 experimentation would allow MyIBD clinicians to

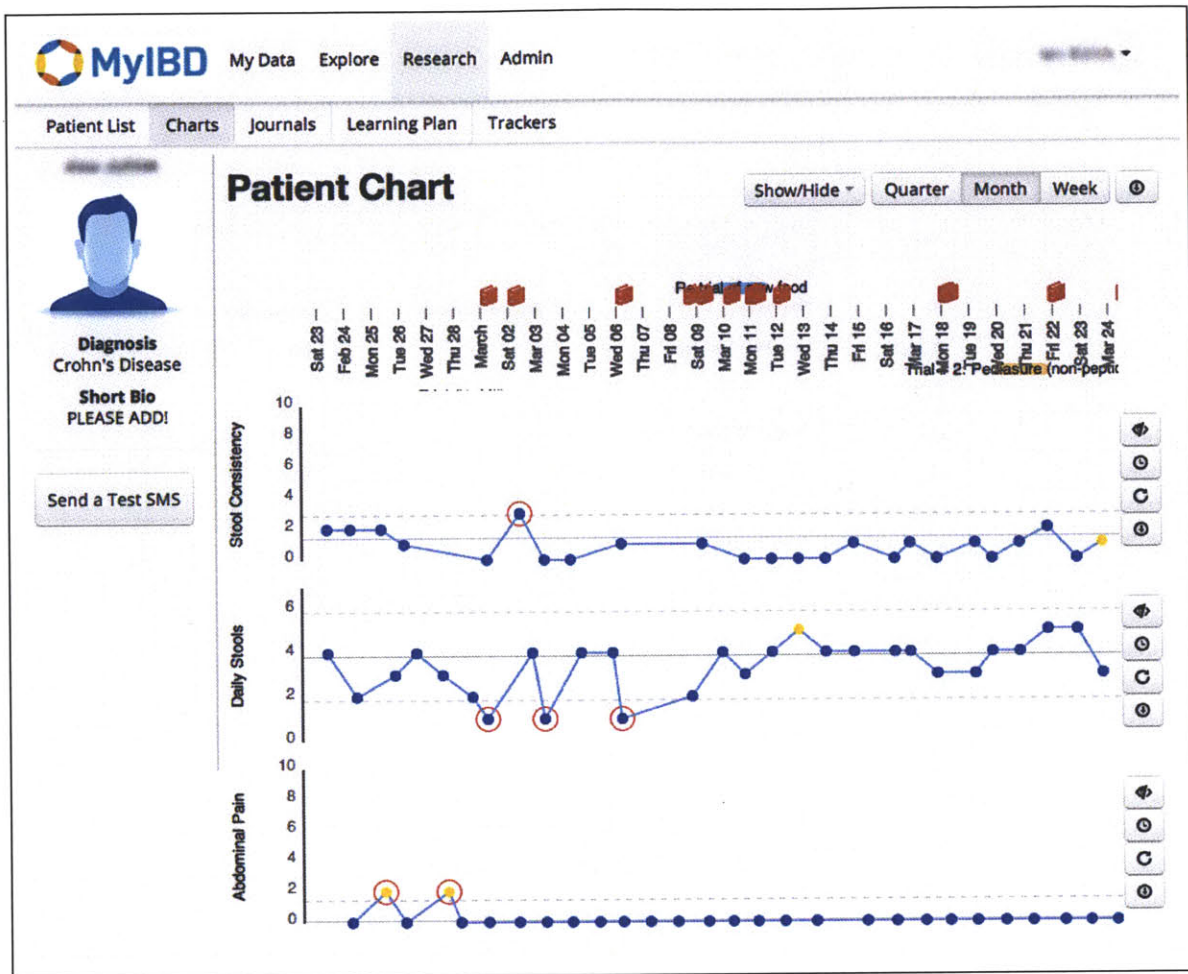


Figure 8-29: MyIBD Clinician Chart Review

more readily embrace and support activities that families are already engaged in, as well as evaluate formal medical treatment for primary disease indicators when the patient is at equipoise with regards to treatment options.

The focus of MyIBD is to facilitate long term optimization of patient health outcomes and quality of life. This contrasts with the design goals of PersonalExperiments, to extract maximal individual and personal inferential power from sequences of lightweight experiments. The contrast between these two views will undoubtedly influence some of the design choices that informed the implementation workflows for Personal Experiments.

The learnings provided by the pre and post MyIBD deployment are documented in Section ???. The use of the structured experiment model and aggregation mechanisms dis-

cussed in Chapter 7 are scheduled to be deployed and evaluated in the trial after feedback from early patients.

## 8.8 Technology Architecture

A brief overview of the technical implementation used for Personal Experiments and My-IBD platform may assist readers who are interested in reproducing this design. The specific choices made were intended to satisfy a number of interrelated goals:

- Enable highly interactive searching and sorting
- Ensure fast page load times, particularly for mobile users
- Don't lose inbound tracking data
- Support a rich hierarchical data model
- Support large sequences of time series data
- Perform offline statistical calculations
- Perform offline fetching of data from 3rd party services
- Ensure HIPPA/HITECH compliance for MyIBD
- Built multiple personalizations of the site from the same code base

The site consists of the two major subsystems shown in Figure 8-30. First, a front-end written in Coffeescript<sup>6</sup> to facilitate the interactive user interface. Second, a backend server was written in Clojure, a mostly functional language for the Java Virtual Machine. The two sub-systems communicate using a “RESTful”<sup>7</sup> application programmer interface (API) that facilitates creation of independent services on top of the data model, such as native smartphone applications.

A subset of the site's pages are rendered directly on the server for convenience, but pages such as the catalog browser, dashboard and timeline that require interactivity are ren-

---

<sup>6</sup>A syntax layer translated directly to native Javascript

<sup>7</sup>REST stands for “representational state transfer” and implies that clients use stateless, primitive HTTP operations to modify one or more data elements on the server.

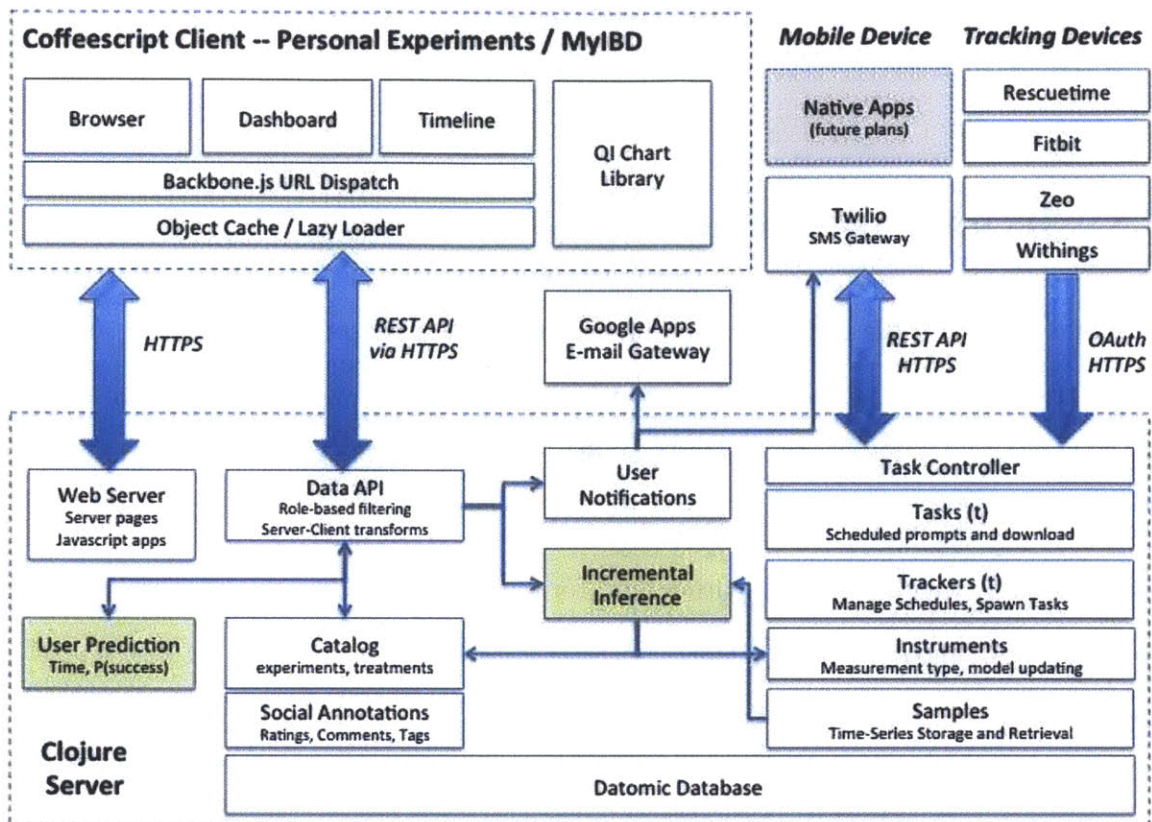


Figure 8-30: Platform Architecture

dered by the server as an embedded javascript application plus inline data to pre-populate the client application’s data model. Subsequent operations are performed entirely on the client side. Data is fetched from the server as needed and side effects are pushed back to the server for persistence.

### 8.8.1 Server Architecture

The server consists of roughly 15,000 lines of Clojure code implementing the subsystems diagramed in Figure 8-30. It exports two major interfaces; a web server and a set of APIs. The supported APIs include a generic REST API for fetching and updating data models, a search API, a set of auto-completion APIs, a template download interface, and charting APIs. The Clojure webnoir framework was used to provide the individual URL handlers,

session handling, and other standard site elements. Data elements were stored in Datomic, a transactional, time-aware database consisting of covered indices over entity-attribute-value-transaction triples.

The server system was built on the principle of statelessness; no critical system state should be required past the duration of a single web request or background task invocation unless it is stored in the database. For example, sessions are stored in a Datomic partition, and the state for the hundreds of active data collection tasks that may be running at any point in time are managed via status applied to database elements.

Trackers and Tasks are two objects tagged for the purposes of ensuring that certain actions are taken on a calendar time basis. Every 60 seconds, one of the servers will run a query on the transactor that finds and marks as 'locked' any tasks that need to be handled. The handler runs to completion and the objects are released. Locks timeout after 5 minutes in case of a node failure.

Every Datomic write results in a propagation of the transaction side effects to all peers. Datomic provides a watcher interface to hook into those writes. This feature is used to ensure that all object statistics are updated incrementally as the state of the system is changed. To avoid coordination, all statistics updates are written to be idempotent, so multiple writes have no effect on state.

Actions that take place over the Data API can trigger user notifications including: new comments to a thread a user has posted on or for an object they are an editor for. The administrator is notified of all registrations, object creation events, etc.

## **8.8.2 Client Architecture**

The client architecture is based on the Model-View-Controller model supported by the Backbone.js library. An extension to Backbone.js called Backbone-Submodels was written to enable the server to send lazily-fetched hierarchical models that are deserialized into models with parent references on the client. The REST API provides a generic API for

creating and updating both ordinary models (no parents) and embedded models (maintain parent links). The templates used to render HTML on the client side are based on a custom handlebars-clj library, that generates handlebars templates that can be evaluated either by the server or the client. These templates can be fetched on demand, or pre-loaded in a tag when the original application page is rendered.

The remainder of the roughly 5,000 lines of coffee script consists of view rendering and event handling code. Model side effects result in the Backbone.js library informing the server of side effects, and perform event-driven re-rendering if the model is updated by the server side in response.

### **8.8.3 Deployment**

Ensuring that all tracker data is received and retained requires a strict uptime guarantee while facilitating frequent feature additions requires a simple, fast upgrade procedure. An operations automation environment based on GitHub and Hosted Chef manages the complexity of maintaining two site installations and testing infrastructure. Deploying on a cloud computing platform, such as Amazon's AWS, adds complexity but also supports the rapid creation of temporary testing facilities running against the main site's database. Using the Chef framework, administration, upgrades, backups and testing of these sites were reduced to a few scripts run on a development workstation. Moreover, the entire infrastructure is stored under source control and trivially reproducible.

A diagram of a deployed and operating site is illustrated in Figure 8-31. The Datomic database is a distributed system consisting of four components: the DynamoDB key-value store, a write-only transactor pair, a memcached layer for caching pages, and a set of application peers that fetch immutable pages from memcache or the data store. Writes are directed by peer libraries to the master transactor which ensures ACID compliance.

Several key characteristics of the architecture facilitate efficient, fault-tolerant operation:

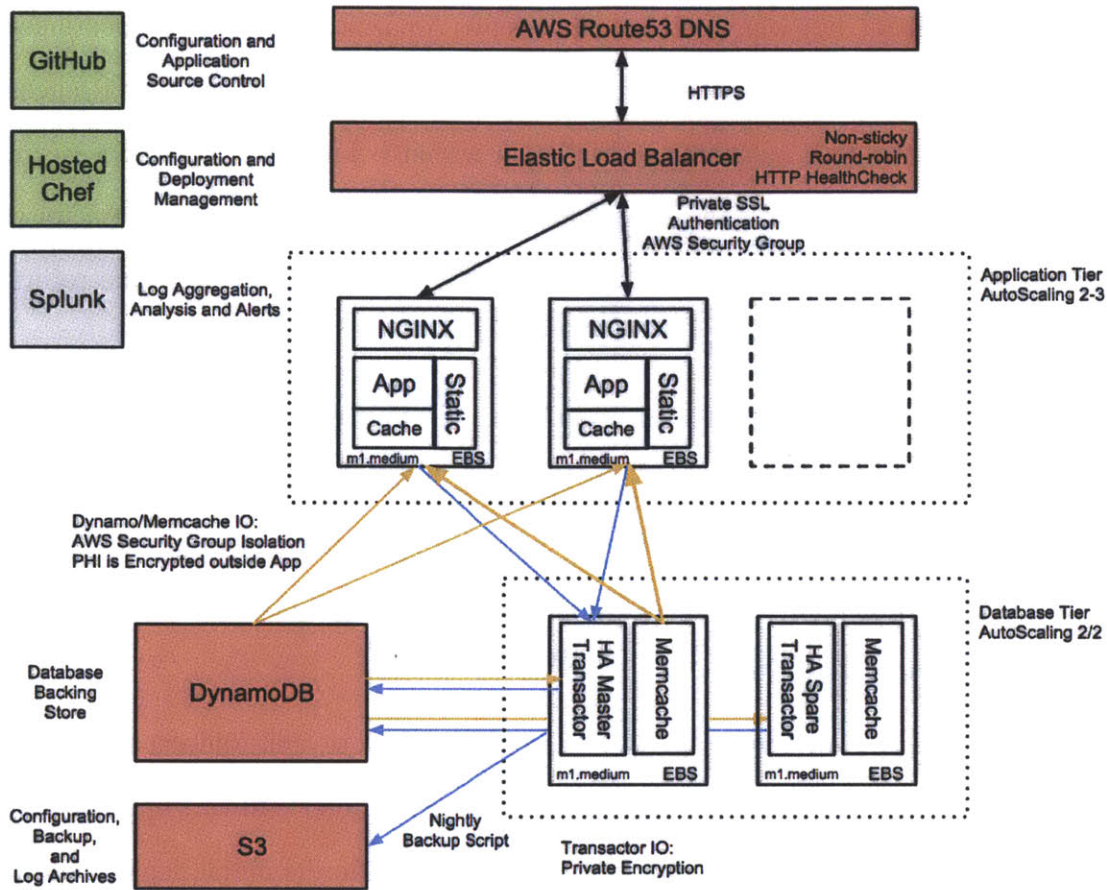


Figure 8-31: Deployment Architecture

- The DynamoDB storage engine is a highly reliable backend data store
- The Datomic persistent database layer maintains a complete, immutable history of all changes to the database, facilitating restoring from any point in time
- The two systems working together simplify deploying, backing up and copying entire databases
- Chef and CloudFormation scripts capture the complete specification of all servers
- Github provides source control to track the history of site configuration and source code
- The database backup plus two source control tags (one for infrastructure and one for the application code base) are all that is needed to reproduce the entire state of the system at any point in time

The web application software is designed to be replicated across multiple servers and in

the MyIBD deployment they were replicated across two Amazon zones to ensure continued operation in case of a zone failure. The Amazon Elastic Load Balancer (ELB) and Route53 DNS system facilitates load balancing and fail over within 1-2 minutes should one of the peers become unavailable. A standby transactor can take over for the master transactor in the case of a machine failure or network partition. Multiple peers also facilitate rolling restarts to ensure near 100% uptime.

The only challenge encountered in ensuring high uptime under this configuration is that the load balancers use DNS to direct traffic, so existing client DNS caches may be stale for a minute before finding a valid peer. This applies to push updates from the Fitbit service and Twilio SMS gateway, leading to some persistent vulnerabilities to data loss that I couldn't justify the engineering time to fix.

The up-front costs of designing the system were paid back several fold via time saved recovering from rare, ongoing events such as hardware faults, cloud hardware upgrades and configuration errors. The site was updated dozens of times after its initial deployment in response to bugs or user requests for new features with a minimum of effort and spinning up test systems to evaluate major upgrades was also simple.

Debugging and hot-patching of the systems is possible via ssh tunnels to connect to a running JVM instance using the NREPL protocol and temporarily enabling sticky ELB sessions to route all traffic for a user to one machine. In an environment with more peers, a testing peer with a sub-URL that is not connected to the ELB but is connected to the live database would be a useful and easy feature to add.

Building a new site from scratch takes about 10 minutes using Chef's *knife* command line tool, including launching machines, configuring the operating system, building the software, configuring the site, installing optional seed data, and warming caches. The only negative to the setup used in this dissertation were difficulties maintaining current security patches on individual servers. This needed to be done manually, or by upgrading the underlying machine image specification and re-launching the site. This delayed patching



represents a short-term security vulnerability that can easily be fixed in future versions.

## 8.8.4 Performance

Performance is ensured through the strategic use of caching throughout the system and minimization of cross-network data calls:

- Certain expensive queries are memoized in the core code
- Datomic peers cache the working data set and recent writes in JVM memory
- Transactor writes are broadcast to all peers in real-time
- Hooks on transactor updates invalidate application caches when needed
- Stale database pages are often stored in a memcache instance (1-2ms)
- DynamoDB uses flash memory for primary storage (10ms)
- Nginx caches all static assets in memory
- All static assets are cached in proxies and/or the browser cache
- All fetched objects are cached in client-side applications

The use of a timestamp prefix prepended to all URLs facilitates caching by proxies or web browsers. The static asset and proxy http server Nginx detects and uses this prefix to set permanent caching headers. This timestamp is incremented each time a site is redeployed. Primary page fetch time is roughly 300ms. Most pages are fully rendered in less than a second except for timeline rendering and initial site loads which can take 2-4 seconds depending on the connection. Performing searches and paging through search results in the explorer webapp takes a few hundred milliseconds under most conditions. These facilities cost about \$300/month for MyIBD due to the replicated servers, and \$70/month for Personal Experiments which runs on a single instance.

## 8.8.5 Lessons Learned

Some of the following lessons may be familiar to experienced software or systems engineers, but they are worth re-iterating for researchers who may not have tackled a project

with these particular constraints. The system is not very large, but this particular set of constraints yielded a surprisingly complicated development effort.

- Rich-client applications, unless you are already an expert web developer, are not a good investment of time for researchers. Instead, for research systems, use server-generated pages with lightweight javascript. For very carefully selected functions, such as visualizations, build standalone client-side logic. This advice may change as technology evolves (see next bullet).
- Javascript is a poor environment for building complex applications, but great for enriching server-generated pages. For applications, look to emerging languages that compile to Javascript such as Clojurescript.
- No matter how tempting, don't use a generic object or key-value API to connect your client and server. This avoids duplication of object state-management object between server and client and simplifies role-based security and debugging. Export a semantic API that captures the logic of your domain at the client-server boundary. You can then capture and reply the entire sequence of high level calls that produced a given error case. (see ThinkRelevance's Pedestal for an alternative strategy to Backbone.js)
- The representation and management of time should be deeply considered up front for any system managing time-series data submitted by users.
  - It helps to have clear boundaries where a certain time base is in force. Timestamps in the database were stored in UTC, any user specified times were all stored as “wall time”, server code and events all assumed the server's local timezone.
  - When data was rendered for the client, it was converted from UTC to the user's timezone.

- When does one day end and another day start for measures that are captured during periods of wakefulness? Some users stay up past midnight.
- Time is particularly tricky if you have users who travel. For example, is the time-series sequence based on the wall time at the point of collection, or the actual global timestamp at which an event occurred?
- Statelessness everywhere is worth the design effort. For example, being able to shut down and restart a server to work around difficult bugs can be very useful.
- Devops tools like Chef and Puppet are worth the investment, but prepared to invest heavily the first time around.
- Debugging time is proportional to reproducibility and inspectability. (See event sourcing)
- Avoid building production systems to generate data for doctoral research projects. The magnitude of effort is rarely justified by the ultimate scientific return. Such projects will make more sense if you are a professor or research scientist and can amortize a series of research projects and/or students over a single platform development effort.

The most challenging aspect of the server implementation tended to fall into the management and coordination of calendar-driven events. The site uses browser detection to identify the user's timezone as well as allowing the user to specify a specific preference, for example when they were traveling. Keeping track of whether a given representation of time was relative to the server's timezone or the user's timezone is something that would have benefited from better abstraction. A mobile app that detected timezones during travel and automatically alerted the server about the user's location would also be invaluable.

Writing all the "business logic" to manage tracking events was also challenging. An SMS request with an inappropriate value, or for a tracker that was just disabled, exposes

many corner cases. Implementing stubs for all the interactive and third party service to facilitate testing would have accelerated working through the various corner cases.

## **Part III**

# **Evaluation and Discussion**



# Chapter 9

## Evaluation

In the introduction I introduce four central questions addressed by this dissertation:

1. Can users design self-experiments?
2. Can users execute and learn from trials of self-experiments?
3. Can we learn from the outcomes of multiple trials?
4. How will this impact the healthcare system?

This chapter answers each of these questions through analysis of external datasets, datasets generated by users of the prototype systems, and targeted user studies. These results demonstrate that individual users are capable of engaging more directly with the tools of science if those tools accommodate their level of activation and the practical constraints that arise in everyday life.

### 9.1 Users Can Design Self-Experiments

The properties of a self-experiment were introduced in detail in Sections 6.2 and 8.2.3. Users need to be able to create instrumentation, document the properties of treatments, and combine these elements together with a specific experimental design. Some of the design parameters that require technical skill include:

- A hypothesis or clear question

- A specific protocol for the intervention (e.g. dosage)
- A quantitative measure for the outcome variable of interest
- Treatment onset and washout times
- Experimental controls
- Randomization

This section presents the results of formative studies as well as examples of user-defined experiments from PersonalExperiments.org to illustrate the elements of experimental design that users understand. For those elements that proved more difficult for users, I demonstrate how the prototype design improved the quality and completeness of the resulting designs.

### **9.1.1 Spontaneous experimentation on social media forums**

I was first inspired to think about user-directed experimentation when I observed examples of formally documented experiments on the TalkPsoriasis forum. These examples use many of the concepts important to creating valid experiments such as protocols, measurement, and timing. Missing from the examples were concepts that support external validity. For example, it is rare to see users propose controls against confounding factors such as confirmation bias, placebo response, or other independent variables.

One of the most intriguing examples of a spontaneously reported experiment was the “Slippery Elm Trial” (Table 9-1, also see full text in Appendix A.2.1).

This example contains a clearly defined hypothesis (leaky gut), a treatment protocol, an outcome measure (photographs), and some exciting preliminary evidence. However, it lacks an explicit experimental control and any documentation of the onset time for the treatment. Unfortunately, the author reported a dental infection that caused a flare-up<sup>1</sup> and stopped posting on the site afterwards.

---

<sup>1</sup>Many people with psoriasis experience flares when they have bacterial or viral infections; often the initial onset of the disease’s skin symptoms comes immediately after an illness.



Concept	User Text
Hypothesis	I have reread all the theories and my analytical mind has assisted me in selecting two causes for this ailment. I believe fully in the leaky gut syndrome however I must also give credence to chronic inflammation as my body seems to exist in this state.
Protocol	5 Slippery Elm caps emptied into water, allowed to sit 5 mins then consumed. Blue berries mixed with 2 Gluten limited not fully free but mostly. Limiting fats except those in flax, fish, and primrose oil. Taking a probiotic supliment Soaking in salt water 1-2 times a week time permitting later in the day so I dont burn. Left salt on my body 12 hours. Avoiding processed foods like the plague.
Outcome	I began this trial July 19th 2011. and have been updating my condition once a week in photos.
Baseline	I am currently 3/4 covered. The P has attacked my fingers and nails for the first time. Both thumbs, index on left hand and pinky on right.
Treatment Result	I awoke this morning with a major over night healing. This morning my body is mostly smooth and flake free. My P is white with small spots of red scattered but limited.

Figure 9-1: Slippery Elm Trial

Was this a spontaneous result, or tantalizing evidence that some patients may benefit from dietary and supplement interventions? The author of this post had a long history of severe psoriasis and attempted treatments without a response of this magnitude. The temporal correlation between the onset of the protocol and the clearing of this user’s skin is strong evidence for the treatment being the causal agent of that change. However, the user has been trying treatments for many years and it may simply be a spontaneous remission that happened to co-occur when they tried this particular treatment. The motivation for introducing controls into trials is to increase internal validity, the confidence that the treatment is responsible for the change in outcome.

What aspects of experiment design do users consider naturally, and which do they need help with?

## 9.1.2 Edison: The Experimenter’s Journal

In 2011, I gave a talk at the Quantified Self conference to try to find other people interested in issues of methodology. The talk was titled “What can I learn from your self experiment”. It proposed that the community engage in a discussion around lightweight, informal ways that QS members could share standardized measures, measurement tools, interventions, and designs. After that talk I met Matthew Cornell, an engineer who had developed a site called “Edison: The Experimenters Journal” to help individuals document, share, and discuss their self-experiments [Cor].

Edison was active from mid-2009 to mid-2012. In mid-2012, Cornell shut down the site due to a lack of time and money to invest in promoting, maintaining, and improving the site. During those three years, it attracted 800 registered users, of whom roughly 115 created 218 public experiments, and generated 1,179 comments.

A compelling feature of the Edison design was a lightly structured format that allowed users to document any ad-hoc experiment. It advocated mindful documentation of what was to be tried and how it would be evaluated. Until its last year of operation, it provided no formal mechanisms for tracking observations or exercising experimental controls.

Matthew Cornell made a copy of Edison’s public experiments available for this analysis (see examples in Appendix A.1. I removed all experiment and comment contributions by the site’s owners.<sup>2</sup> The resulting dataset provides a good foundation for assessing how a self-selected population of motivated users engages with the concept of experimentation.

### Experiment Schema

The data model of the site consisted of an experiment defined by:

1. **Category** The category of the experiment
2. **Title** The name of the experiment

---

<sup>2</sup>These two accounts contributed 92 and 38 experiments and another 1200 comments not included in the above totals.

- 3. **Description** Description of the experiment
- 4. **Measurement** What will be measured to evaluate the success of the experiment?
- 5. **Completion Test** What is the stopping condition defined by the experiment?
- 6. **Status** Running, Completed, or Ongoing

Informally, the experiments fell into two major classes. The first are best described as a public commitment to a specific task, for example:

Field	Text
Title	Getting rid of shin splints
Description	I will keep a diary to track pain, treatments and prevention exercises.
Completion	It will be a success if I have no pain on 14 consecutive days.
Journey	I will enjoy running much more when I am injury-free.

The second more closely resembled the kind of experiment described here, for example:

Field	Text
Title	Effects of drinking milk on night sweats
Description	<p>I worked out a while back that if I drink a lot of milk just before bed I have night sweats after 5 or 6 hours of sleep. A fair bit of tracking went into working this out!</p> <p>I'm interested to know:</p> <ol style="list-style-type: none"> <li>1. Where the milk threshold lies in ml before I start to experience sweats</li> <li>2. The effect of the amount drunk on the severity of the sweats</li> <li>3. Does it make a difference whether I drink the milk alone or with something else (i.e. breakfast cereal)</li> <li>4. The time to onset from drinking to night sweats</li> <li>5. Whether there is the same effect with unhomogenised milk as there is with homogenised milk.</li> </ol>
Completion	I will have answered my 5 questions.

A manual review of the experiments identified 56 experiments that could be reclassified from “Other” into the existing taxonomy. Figure 9-2 provides the distribution of experiment

<b>Category</b>	<b>Original Count</b>	<b>Reclassified Count</b>
Other	120	66
Health/Medicine	23	42
Diet	18	18
Work	14	31
Fitness/Exercise/Sports	9	18
Hobbies/Recreation	7	7
Home/Garden	3	7
Relationships/Dating/Sex	3	3

Figure 9-2: Edison Experiment Categories

types in Edison.

To align with the health-oriented experiments described in this dissertation, experiments from the Health/Medicine, Fitness/Exercise/Sports, and Diet categories were coded. The coding classifies experiments according to which of the key experimental parameters were generated spontaneously by Edison users.

Experiments that clearly identified an intervention and an anticipated outcome were coded as containing a hypothesis. Users did not need to speculate as to the causal mechanism, in keeping with the simplified hypothesis space model introduced in Chapter 6. Quantitative outcomes imply that either object criteria or quantitative subjective criteria were employed in the experiment. A controlled experiment implies there is a baseline period or some other clear basis for comparing the measured results. Quantitative measures like weight or number of handstands have an implicit point of comparison, but were excluded if not made explicit in the description. Onset time was coded if the description clearly identified a timeframe over which the treatment was anticipated to work, implying an awareness of time to effect. The rest of the parameters should be self-explanatory. The results of the coding analysis are summarized in figure ???. For additional detailed context, a selection of 10 example experiments from the Edison site with their coding are included in appendix Section A.1.

<b>Code</b>	<b>Count</b>	<b>Percentage</b>
Hypothesis/question?	55	71%
Intervention protocol	73	93%
Quantitative outcome?	42	54%
Controlled?	21	27%
Onset time?	40	51%
Washout time?	4	5%
Covariates	12	15%
Randomized?	2	2.5%

Figure 9-3: Coded Edison Experiments

The analysis suggests that the site encouraged users to pick quantitative measures and experiment durations sufficient to accommodate the onset period and that users were able to reason about these issues. Only a few users articulated any concept of formal experimental controls or washout times. The few cases that illustrated controls, randomization, and washout were correlation studies with comparisons performed against independent observations of interventions with immediate effect and no carryover.

The cases with a quantifiable outcomes rarely had clear controls, but all had at least an implicit baseline reference, although that baseline was rarely measured. The designs on Edison primarily resemble the interrupted time series design (see Section 2.5).

In addition to the coding, I observed that users regularly refer to external sources for definition of interventions and/or experimental design. The ability to refer to external resources is an important capability of an experiment platform. Because only owners can edit the body of experiments on Personal Experiments, I introduced a reference annotation allowing any user to attach a URL to a catalog element.

### 9.1.3 Experiment Design Survey

During the design of Personal Experiments, I performed a small formative study to assess whether providing a background tutorial and explicit prompts would help ordinary users document experimental controls, treatment behaviors such as onset/washout times, and

<b>Code</b>	<b>Count</b>	<b>Percentage</b>	<b>Edison Percentage</b>
Hypothesis/question?	5	100%	71%
Intervention protocol	5	100%	93%
Quantitative outcome?	5	100%	54%
Controlled?	3	60%	27%
Onset time?	4	80%	51%
Washout time?	2	40%	5%
Covariates	4	80%	15%
Randomized?	1	20%	2.5%

Figure 9-4: Experiment Design Survey compared to Edison

important covariates that were not typically provided by Edison users.

The study invited users interested in experimenting to design an experiment using a free-text form after reading 30 minutes of background material and three sample experiments. Users were asked to design two experiments: one for a pre-defined treatment (glycerin and witch Hazel) and a second for an intervention of their choice. A short follow-up survey assessed what they were able to retain from the introductory material.

The background material was not accessible for most people. A number of people who registered and started the process did not complete the study. The Edison coding methodology was applied to the responses of the five users who did complete the study:

Despite the small sample, the education and prompting clearly influenced the structure of the designed experiments. The hypothesis, protocol, and quantitative measures were consistently provided by all participants. The biggest change between the Edison group and the study group was an increased prevalence of covariates.

Although there was an improvement in the percentage of users who discussed controls and onset timing, washout times and randomization continued to be a challenging concept for users. Only one study provided a truly randomized design, and that was by a college student who had recently taken a course on methodology.

<b>Title</b>	<b>Assessing the effect of 500mg of L-Tryptophan before bed time on hours of deep sleep</b>
<b>Hypothesis</b>	It's a precursor for serotonin which is converted into melatonin. There may be other mechanism. I'll look into this more.
<b>Outcome</b>	Deep Sleep
<b>Covariates</b>	None
<b>Design</b>	BA – 6 nights on L-Tryptophan, 5 nights off
<b>Onset/washout</b>	1 hour to take effect, washout unknown (89 hours?)
<b>Treatment Description</b>	A protein that is a precursor to melatonin via serotonin
<b>Side Effects</b>	None known.
<b>Protocol</b>	Take 500mg L-Tryptophan 15 minutes before bedtime. The goal for bed time is 12.15am. The L-Tryptophan should be taken at 12am.

Table 9.1: User Experiment of L-Tryptophan

### 9.1.4 Designs from Personal Experiments

When Personal Experiments was released to the general public, I helped users improve their designs interactively. For example, figure B.2.1 details one of the early experiments, a two-week test of L-Tryptophan. This user provided all the parameters of a design, except for a washout period where they indicated they had no idea of the appropriate washout period.

Other experiments included:<sup>3</sup>

- Increasing exercise to improve sleep (9 users)
- Daily exercise to improve my energy level
- Tea tree oil for *Tinea Versicolor*
- Going to bed early will improve my mood
- Improve sleep with melatonin
- Increased sleep will lead to better focus

There were a few common confusions demonstrated by users. Most users approached the site from a consumer's perspective, and like Edison users, confused the concept of an experiment (the template of a trial) from their own trial (an instantiation of the template).

---

<sup>3</sup>More user-designed experiments can be seen in Section B.2

The consequence of this misunderstanding is primarily a repetition of the same basic experiment in the catalog, the very proliferation the design of the experiment model was intended to reduce. A visualization that more directly focuses user attention on existing experiments, rather than asking them to search for one before creating it, will be critical for avoiding unnecessary duplication of experiments. However, if experiments are similar, merging experiments for purposes of aggregation remains feasible.

### **9.1.5 Discussion**

The evidence suggests that most self-selected users are capable of thinking effectively about simple hypotheses, quantitative measurements, protocols, and comparisons to a baseline. However, concepts of experimental control such as randomization or treatment-withdrawal pairings, even when well-motivated and explicitly prompted, appear difficult for most users, although they will estimate or guess if guided to do so. The ability of the framework to estimate carryover effects from trials and surveillance will be crucial to effectively scaling the model.

Instead of asking users to design the protocol, the system is likely to be much more robust if it asks users to characterize the constraints that would lead to the selection of a protocol. In particular, randomization and the number of alternating treatment-baseline pairs are selected to balance confounding influences. Users are capable of understanding the concept of confounding and so may be much better at characterizing the frequency and magnitude of confounding factors than designing a trial. See Section 10.2.5 for a further discussion of this idea.

## **9.2 Users can Execute and Learn from Self-Experiments**

The second main question of the dissertation involves the experiences of users who engage in experimentation. We want to know how well they follow the experimental protocol and



whether they are able to learn from their individual outcomes. Several different workflows were used to evaluate this question. This includes the manual n-of-1 trials observed within the C3N Project, the MyIBD prototype deployed within the ImproveCareNow network, and the PersonalExperiments.org prototype site. This section emphasizes the learning from a user study run on PersonalExperiments.org.

### **9.2.1 C3N/ImproveCareNow N-of-1 Experience**

Early in the development of the research, the n-of-1 team at Cincinnati engaged 10 test subjects in months-long trials of interventions such as antibiotics, Maalox, and probiotic blends. These subjects were exposed to a day-to-day experience similar to that of future users of MyIBD or Personal Experiments. They received staff-created SMS prompts on a pre-defined schedule, responded to them from their cellular phones, and staff personnel transcribed the responses into an Excel file. Excel files and copies of charts in PowerPoint slides were used to share and review the information.

The following story is from a case study illustrating the value of tracking and experimentation as part of a larger continuum of care, rather than through isolated techniques.

One IBD patient family gave pharmaceutical grade probiotics to their child in an attempt to reduce their nighttime stooling frequency (6-9 times a night). Getting up many times each night had a significant effect on the patient's quality of life. Unfortunately, there is a small amount of mostly negative results on the use of probiotics and antibiotics treating IBD, but nothing was found in a literature search specifically for nighttime stool frequency. The probiotics had little effect, but during the trial, the patient was treated with amoxicillin, a systemic antibiotic, for a sinus infection. During treatment, the patient's nighttime stooling stopped entirely, a fact that she did not recall until she and her clinician reviewed the data together. The clinician then hypothesized that the same effect could be realized with rifaximin, a non-absorbable antibiotic that could

be prescribed long term, and they planned a new trial to test it. Unfortunately, that patient had complications from their primary disease that prevented evaluating the Rifaximin, but that design is now available for other patients with concerns about stool frequency.

Adherence to SMS tracking for these experiments was over 90%. No significant failures to complete the intervention schedule were reported. Similar adherence numbers were observed with the subsequent deployment of MyIBD. More details of the MyIBD deployment are reported in Section 9.4 below.

## 9.2.2 Personal Experiments for Self-Tracking

The uses of the Personal Experiments site for self-tracking were highly encouraging. Users found it easy to register, select a set of measures, and report data over SMS. In the 4 months after the site was advertised on several social media sites, it received 309 registrations yielding 217 active trackers and 35 trials. The following tables illustrate the popular tracking measures and trial outcomes preferred by users, including the contributions of users in the user study reported in the next section.

<b>Variable</b>	<b>Count</b>	<b>Variable</b>	<b>Count</b>
Total Sleep	25	Fitbit Activity Score	7
Steps	21	Computer Productivity	7
Deep Sleep	18	Time Awake	7
Time to sleep	17	Total Bed Time	5
Sleep Quality	13	Computer Work Efficiency	5
Weight	11	Mood	4
Nightly Awakenings	8	Distance	4
Deep Sleep Ratio	8	Total Calories	3
Total time on computers	7	Sleep Efficiency	3

<b>Treatment</b>	<b>Count</b>	<b>Treatment</b>	<b>Count</b>
Increase in Exercise	13	Tea Tree Oil	1
Self-Control Application	5	Pomodoro Technique	1
flux Application	4	Melatonin	1
Adherence to Medications	3	Glycerin and Witch Hazel	1
White Noise	2	Buddy Up	1
L-Tryptophan	2		

### 9.2.3 Personal Experiments User Study

This section presents the results of a user study to evaluate the viability of the Aggregated Self-Experiments model. I recruited 22 participants interested in experiments relating to sleep or productivity. The users were split into two cohorts: a Treatment group and a Control group. Recruitment was performed via e-mail sent to Quantified Self mailing lists and social media forums, as well as to users who registered on Personal Experiments. The Jawbone UP device was provided to all users as an incentive to participate, as well as used to track various sleep variables. The only restriction on participation was that subjects were adults in good health.

The Treatment group was presented with a list of five productivity experiments and five sleep experiments from the site catalog. They were asked to choose one and execute it without any additional input from the investigator (other than site technical support). The Control subjects were presented with the same set of treatments as the Treatment group and asked to perform them on their own without use of the site or the experimental protocols documented there. The treatments and experiments used in the study are documented in Section B.1.<sup>4</sup>

After two weeks, all the control subjects were interviewed about what they tried, what

---

<sup>4</sup>It should be noted that to control the total time of the study and evaluate user reactions to the process of experimentation rather than focusing primarily on outcomes, the protocols for these treatments were shortened to one week phases that could only detect large changes.

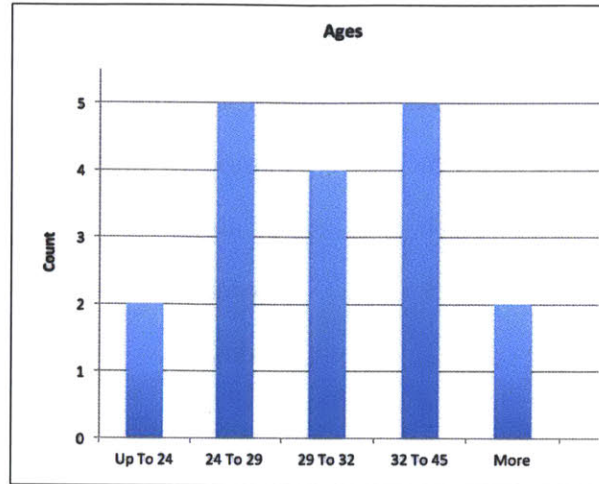


Figure 9-5: Population Ages

they learned, and what their problems were. They were then given an opportunity to use the site to repeat the experiment, or to try a new one with some input and guidance from me. This interrupted time-series design provided a control arm for the treatment group, as well as a matched comparison between standard care (or no tools support) and use of the Personal Experiments website. The characteristics of the study population, results of the site experiments, and interviews are detailed below.

## Population

The population consisted of people who were “QS-curious,” meaning they were interested in the Quantified Self community and/or had tried to track some aspect of their health in the past, but few had used self-tracking data to directly improve their lives. Four of the twenty-two participants were female. Figures 9-5 through 9-8 summarize additional characteristics of the user population including occupation and degree of prior self-tracking experience.

## Interviews

I engaged in 40-60 minute phone interviews of twenty of the twenty-two subjects.<sup>5</sup> The highlights from the interviews are:

<sup>5</sup>The remaining two subjects had logistical problems or were unresponsive to my outreach.

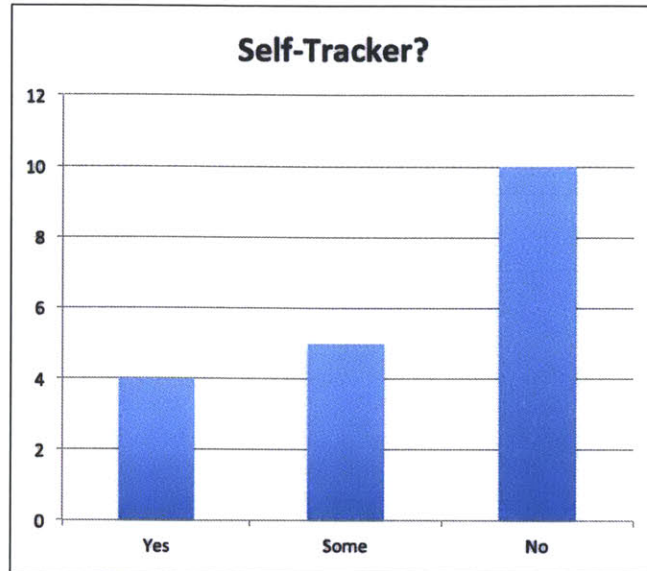


Figure 9-6: Prior Self-Trackers?

- 80% of interviewed users explicitly stated that the experimental context changed their view of self-tracking,
- Three users said that the treatments they chose “changed their life” and wouldn’t have occurred without the external impetus to experiment,
- 85% of users engaged in discussions about their concrete goals, and
- 95% of interviewed users wanted to continue working with Personal Experiments after the study.

Because most of the population were not avid self-trackers, the strong response to the concept experimentation suggests that bringing a concrete, finite goal to the self-tracking process greatly improves people’s interest in engaging in and adhering to the tracking necessary to power the experiment. One former self-tracker says: “I have two years of Fitbit data, but I have no idea what to do with it.” This sentiment was mirrored by other subjects who had previously tried to collect data. Of the study participants, only two had experienced significant learning from prior self-tracking experience.

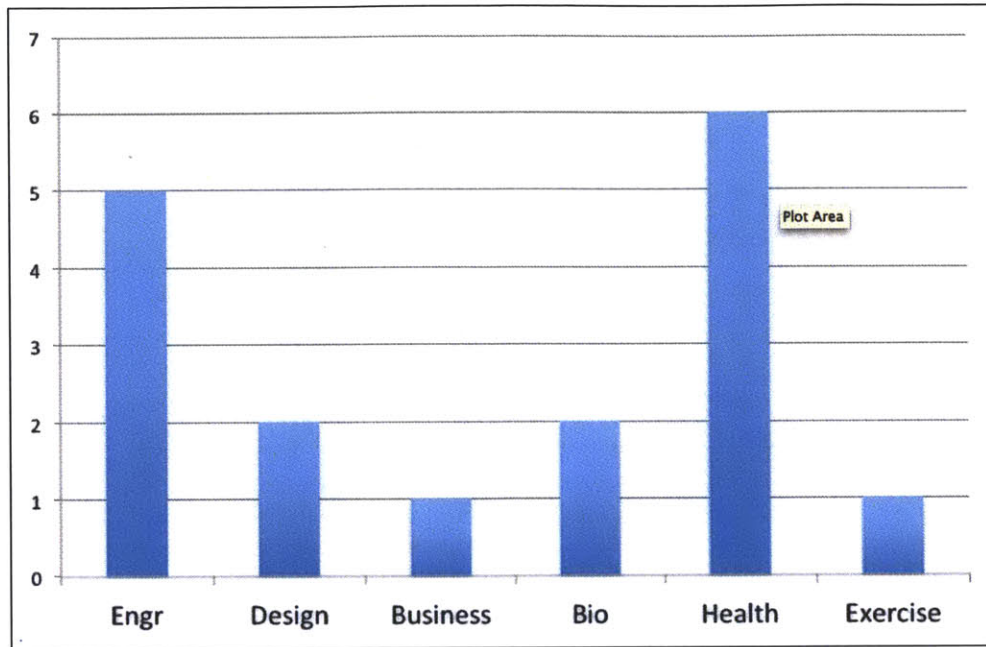


Figure 9-7: Population Occupations

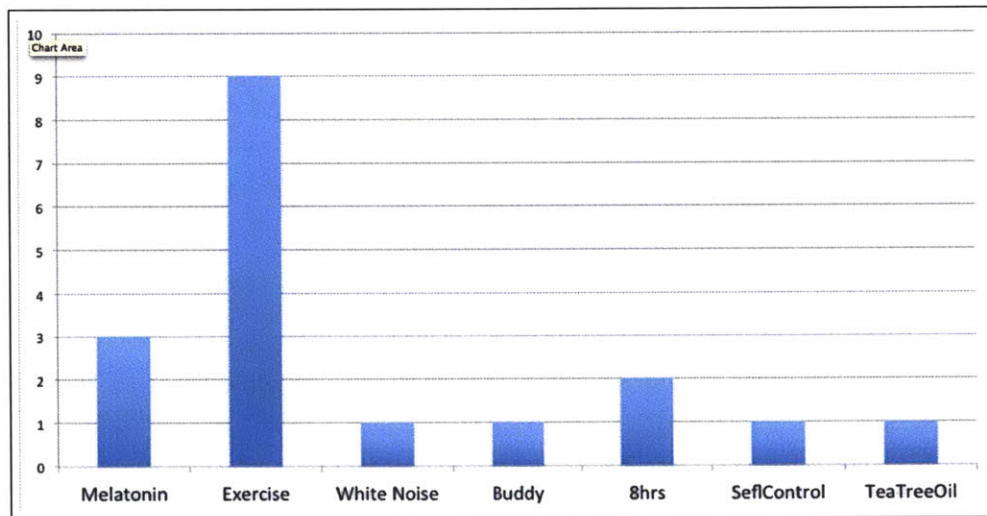


Figure 9-8: Selected Interventions

The fact that an experiment is of finite duration was crucial for most users. If the incentive is sufficient, anything can be done for a few weeks. Users agreed that a handful of weeks is the range of time they are willing to spend on an experiment; an experiment of many months would require a much stronger incentive, such as that of a long-time psoriasis sufferer.

When users tried a treatment, but saw no tangible difference between their treatment and control data, they were more likely to be confident that the treatment didn't work than users who tried a treatment without using the site.

The users who experienced the large effect sizes were by chance the two control subjects who opted to try 8 hours of sleep a night during their two week exploratory period. Both users were among those who said that the experience changed their life. It was the experimental context more than the tools that enabled them to try the short term intervention. Part of what the treatment description communicated is that the experiment will establish how good you could feel, not that you need to sleep 8 hours forever. The follow-up for both users is investigating how to improve what sleep they do get, and to discover what amount of sleep they should target that is less than 8 hours but more than they were getting previously.

A third control subject investigated the impact of exercise on total sleep, but independently decided to look into his ratio of deep to shallow sleep as measured by the Jawbone UP device. He was motivated to do this by a report from Jawbone that he was in the bottom 10% of all users. When he increased his exercise, he observed that he went from 15% up to about a 35% deep sleep ratio and that was associated with a perception of better energy the day following. He wanted to repeat this experiment on Personal Experiments during the follow-up phase, but hadn't started it by the time of this writing.

The user who didn't opt to continue using the site said that they were uncomfortable using the site because it kept a history of their failures, and being forced to visualize or review past failures was disturbing.

The following case studies illustrate specific experiences.

### **Trial of the Impact of Exercise on Sleep**

Figure 9-9 is an example of a user 23 year old male with a desk job demonstrating a weak but significant shift in the mean amount of sleep during the treatment period. The second

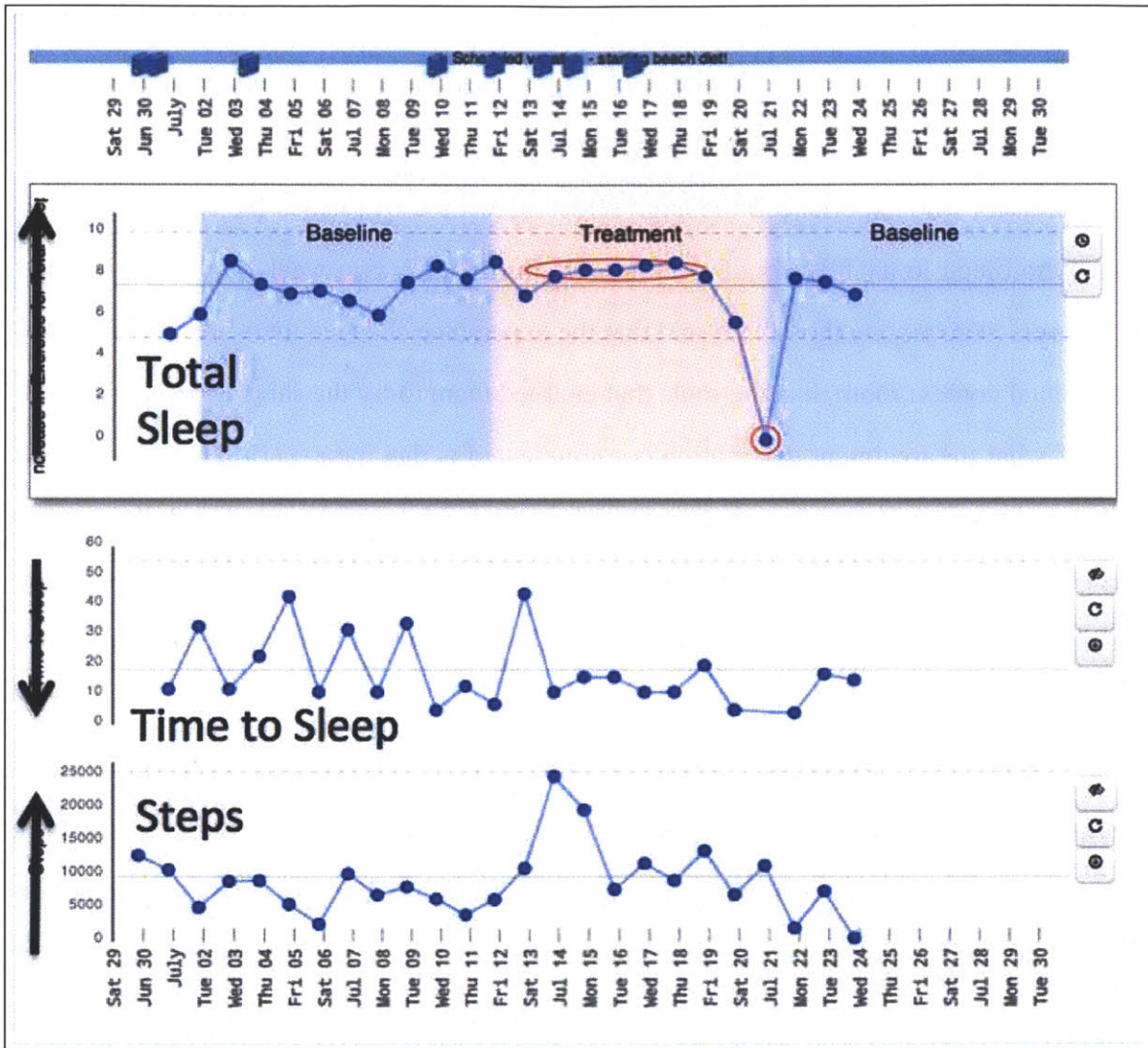


Figure 9-9: Trial of Increase Exercise to Increased Total Sleep

run chart in the figure is interesting in that the variance in time to sleep was decreased during the treatment period. Also note that the third chart, total steps, failed to document a large change in activity level as the exercise took place in a gym where a wrist-mounted accelerometer is a poor measure of exertion.

Despite the outlier (an all-nighter) on the last day of the treatment phase, the trial itself was successful as measured by the decision rule that fired when there was a run of 5 points above the mean measure. The effect size, however, would be considered to be small; the effect size was only about 15 minutes.



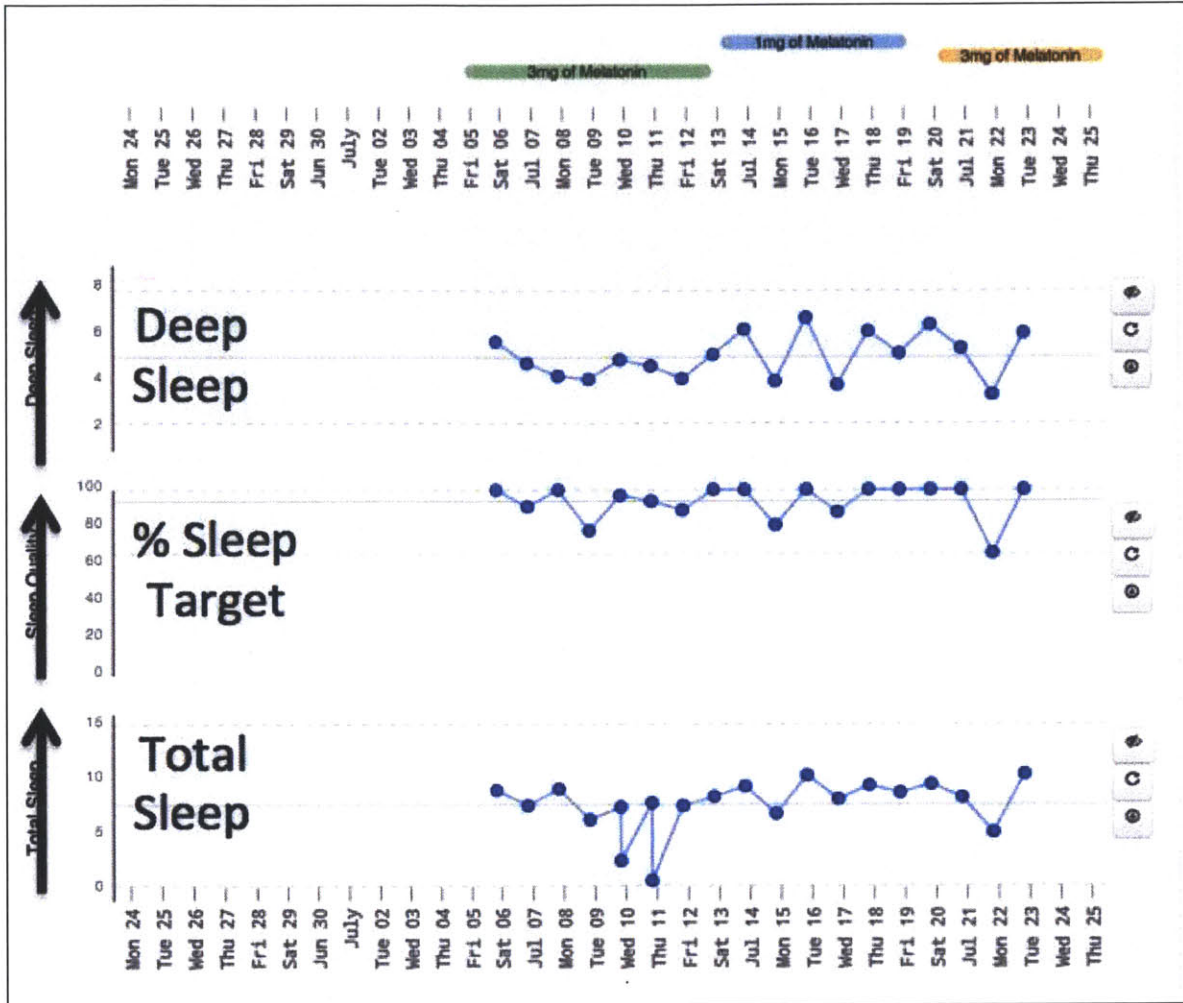


Figure 9-10: Trial of Melatonin Dose impact on Deep Sleep

One final note is that the outliers for this user occurred on weekends. The ability to take into account factors like weekends and travel on experiments would be valuable for future versions.

### Trial of Melatonin Dose-Response

One study subject was a 69 year old male with Huntington’s disease who has spent many years optimizing his sleep patterns. Figure 9-10 shows how he used the journaling feature to highlight periods of time where he was taking different doses of Melatonin to see if there was a visual difference between 1mg and the 3mg doses he had been taking previously.

This is a great example of how users will use general facilities for their own purposes,

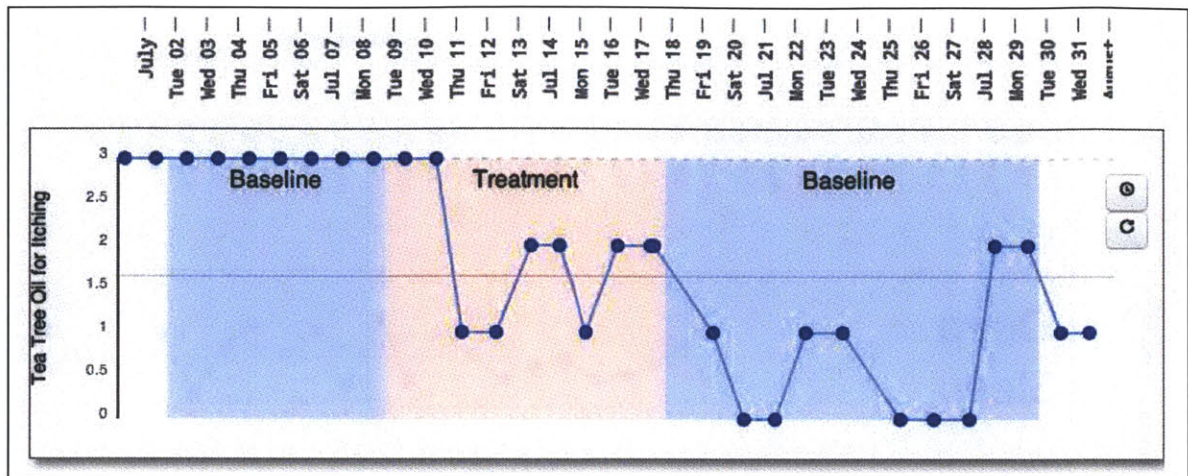


Figure 9-11: Use of Tea Tree Oil to Reduce Itching from Tinea Versicolor

as well as a two-armed trial that compares two treatments. This is also an ad-hoc dose-response study, although it doesn't provide enough information to fit a dose-response curve.

### Trial of Tea Tree Oil

Several online resources say that Tea Tree Oil is a good treatment for the fungal infection Tinea Versicolor. This is one of the most common fungal infections. It causes a rash and in some patients a highly unpleasant itching. There is evidence that the oil is an effective anti-fungal and anti-microbial, but no specific evidence that it treats Tinea Versicolor. One subject, a 45 year old medical professional, asked to perform this specific experiment to see whether the oil reduced the itchiness, the result of which is documented in Figure 9-11.

Two observations from this case are relevant. The first is that the carryover of the treatment (which kills the fungus) is quite long. The second is that the itching does eventually return. This raises the question of whether a longer treatment would imply a cure by killing all the fungus and how the site should accommodate irreversible treatments (for example, multiple trials when symptoms re-occur). This infection is often seasonable, brought on by humid weather, which might provide further opportunities to replicate the experiment in the same subject over time.

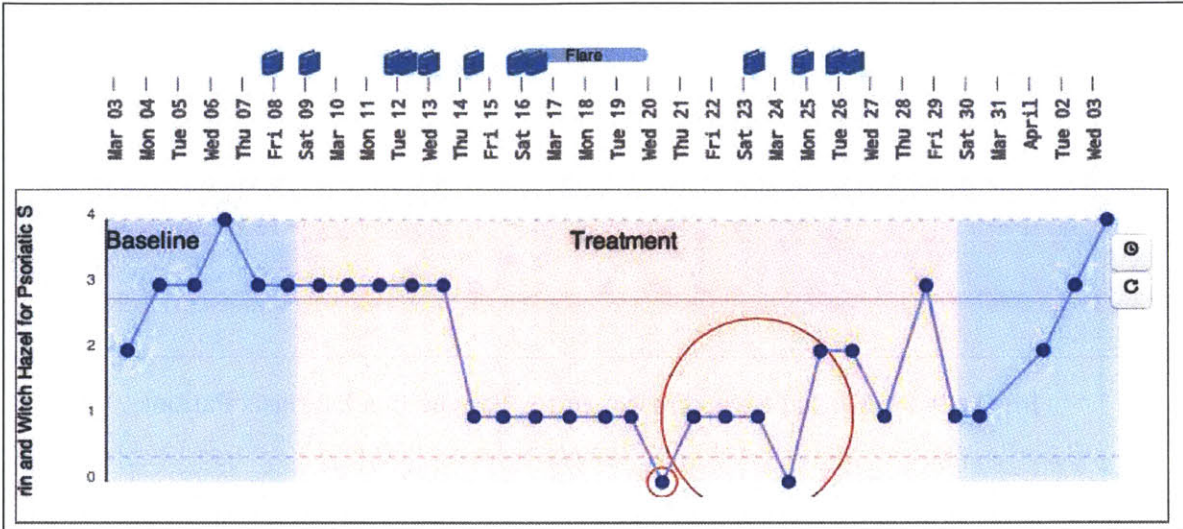


Figure 9-12: Glycerin and Witch Hazel for Psoriatic Scaling

### 9.2.4 Uses of Personal Experiments in the wild

In addition to the user study, there were several compelling examples of natural uses of the site worth mentioning, one of which provided a basis for Alice’s studies in Chapter 5.

#### Trial of Glycerin and Witch Hazel

Figure 9-12 shows the trial chart for an experiment evaluating the use of glycerin and witch hazel on psoriatic scaling. It clearly establishes a large effect and demonstrates the firing of both control rules, out of bounds and a run under the mean, indicating that the treatment phase showed a significant shift in the mean.

As with the tea tree oil example above, this effect was have been obvious to the user even without Personal Experiments. However, the user has a clear record of success to refer to that complements their subjective memory of the trial, and other users can benefit from the documentation of effect size, onset time, and washout time.

#### Trial of Dietary Supplements

One of the treatments recommended to Alice in Chapter 5 is a herbal supplement users say will help with her fatigue. The effects of the supplement shown in Figure 9-13 are less

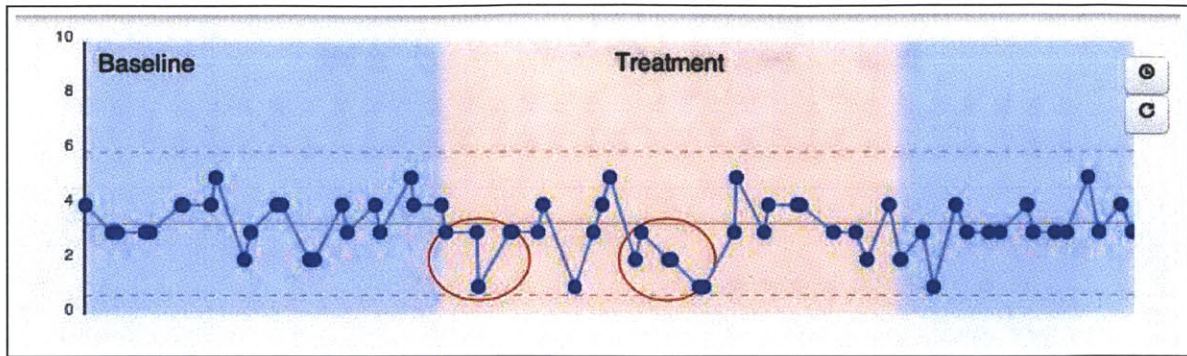


Figure 9-13: Herbal Liver Treatment for Fatigue in a Psoriasis Patient

powerful and more difficult to identify, given the confounding factors present during the study.

The primary outcome of the trial was fatigue reported on a 1-10 metric scale. The control limits indicate that the typical magnitude of fatigue was between 1 and 6 where values below 1 and over 6 would fall at the 1% probability level. During the trial there were two periods of time when the run of points fell below the mean, an event unlikely to happen by chance. The user also noted in their journals that they experienced clear triggers causing flares, the two spikes that are visible during the treatment period.

## 9.2.5 Discussion

The Personal Experiments study shows that with appropriate context and support, most people can run experiments with high levels of adherence to the necessary tracking and treatment schedules.

There are several features in the current design that need to be improved to create a version of Personal Experiments with broad appeal. The site has a relatively high friction to adoption, easily dealt with via some hand-holding, but that would benefit from a careful re-design that more incrementally engages users in tracking. Second, because the design and selection of experiments appears to be challenging for users, even those with incentives, the site needs to find ways to simplify the specification of an experiment while remaining within the data architecture laid out by Chapter 6. Finally, as mentioned above, the catalog

layout function is still inadequate to guide people intuitively through the experiment design function, helping them move from consumer to producer. The prospects for improving these particular features are discussed further in Chapter 10.

### 9.3 Learning from the outcomes of multiple trials

It will be some years before sufficient masses of single-subject data are generated on platforms like Personal Experiments or MyIBD to evaluate the aggregation capabilities of the framework on real-world data. In the absence of large-scale field experience, the commitments of the framework are validated by extrapolation from the behavior of users engaged in tracking and trials together with the assumptions laid out in Chapter 7.

The key mechanisms proposed in Section 7.5 included:

- **Estimating Variance** - Using the results of tracking data and trials of specific measures to optimize sample sizes in future trials
- **Parameterizing a Design** - Select a the parameters of a trial that fits what we know about the actual measure and treatment dynamics
- **Improving a Design** - Learning about the characteristics of measures and confounding to design better trials
- **Predicting Population Outcomes** - Using the outcomes of prior experiments to estimate success prevalence and effect size
- **Recommending Experiments** - Ordering user experiments to optimize the search for viable treatments

The last two are straightforward applications of theory that require much larger populations to evaluate empirically, but we can assess the potential optimization of the first three by evaluating the data produced in the user study.

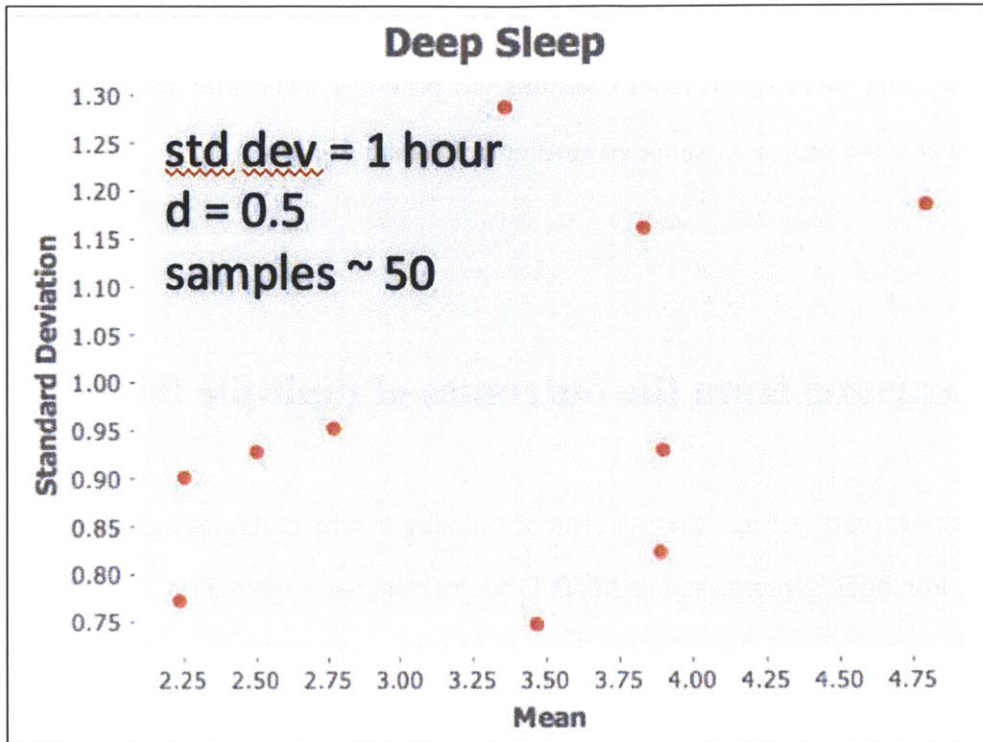


Figure 9-14: Deep sleep has a typical Cohen's  $d = 0.5$

### 9.3.1 Analysis of Variance

Section 7.2.3 makes a case for estimating the variance of a measure to optimize the sample size estimate for a design. What optimization opportunities exist in real world datasets? Figures 9-14 and 9-15 illustrate the standard deviation  $\sigma$  versus the population mean  $\mu$  for two measurements of sleep from the Jawbone UP device.

Deep sleep measures periods of sleep characterized by generally less accelerometer-measured activity than in light sleep and awakening periods. Deep sleep has little correlation to the formal phases of sleep measured by electrical activity. However, it is a good approximation of how disturbed an evening of sleep might be (e.g. restless legs, apnea, environmental disturbances, etc.). Improving the amount of hours in deep sleep is a good goal for users looking to improve the quality, not just quantity, of their sleep.

These two diagrams illustrate how the choice of a measure can influence the sample size of a trial. A measure with a lower variance will require smaller sample sizes. All things

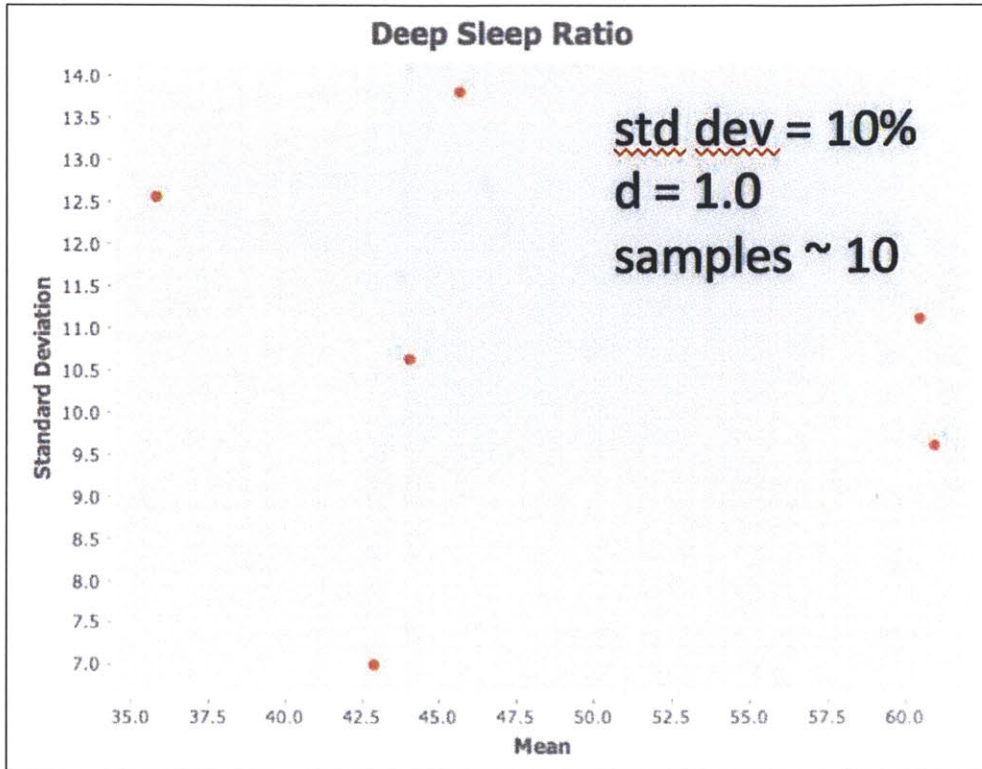


Figure 9-15: Deep sleep ratio has a typical Cohen's  $d = 1.0$

being equal, users should select measures that have lower variance to achieve more efficient trials of the same underlying phenomenon. Here, the ratio of deep sleep normalizes over the high variance in total sleep hours experienced by most users (shown in Figure 9-16), resulting in a more sensitive measure of sleep quality and lower required sample sizes.

### 9.3.2 Optimizing Sample Size

Figure 9-17 illustrates a conventional sample size calculation for a two-tailed t-test of the difference of two gaussian distributions, the conventional model most resembling the model presented in Chapter 7. This is a computationally cheap way to illustrate the impact of minimal effect size and confidence on the required sample size.

This chart is generated by the equation:

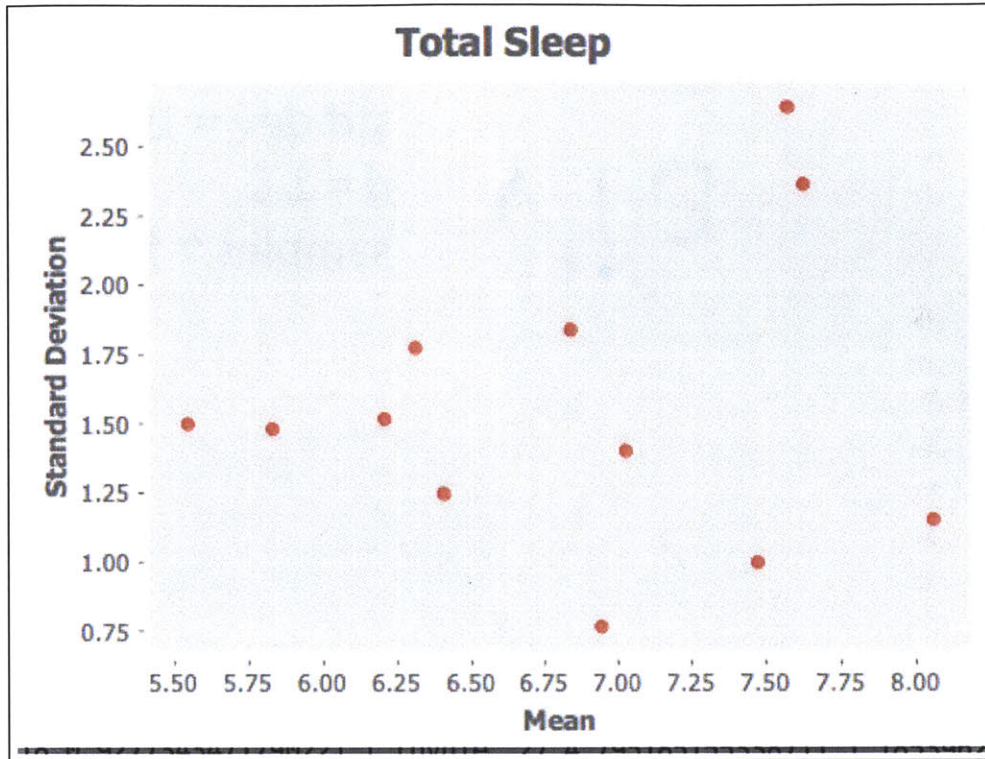


Figure 9-16: Study Population Total Sleep

$$n \geq \frac{2\sigma^2(Z_\beta + Z_{\alpha/2})^2}{\Delta\mu^2} \quad (9.1)$$

Unfortunately, these equations assume that sigma is a fixed quantity, when in fact there is a distribution over sigma that needs to be accounted for. Power calculations that take this extra uncertainty into account will result in higher sample sizes, as illustrated by the simulated sample sizes in Figure 9-18.

For  $d = 1.0$ , the conventional sample size calculation yields a sample size of 10 (for  $\eta = 0.8$ , or a one-tailed false-positive rate of 10%) versus the fully uninformed simulated model, which yields a sample size of 22. The simulated sample size also provides a curve<sup>6</sup> for the model with a tight prior on the effect size being equal to the value of the trial with a tight estimate of the variance. We see at  $d = 1.0$  a 30% reduction in sample size. The value

<sup>6</sup>The curve is not continuous due to the search procedure used to find the optimal sample size at each effect size.



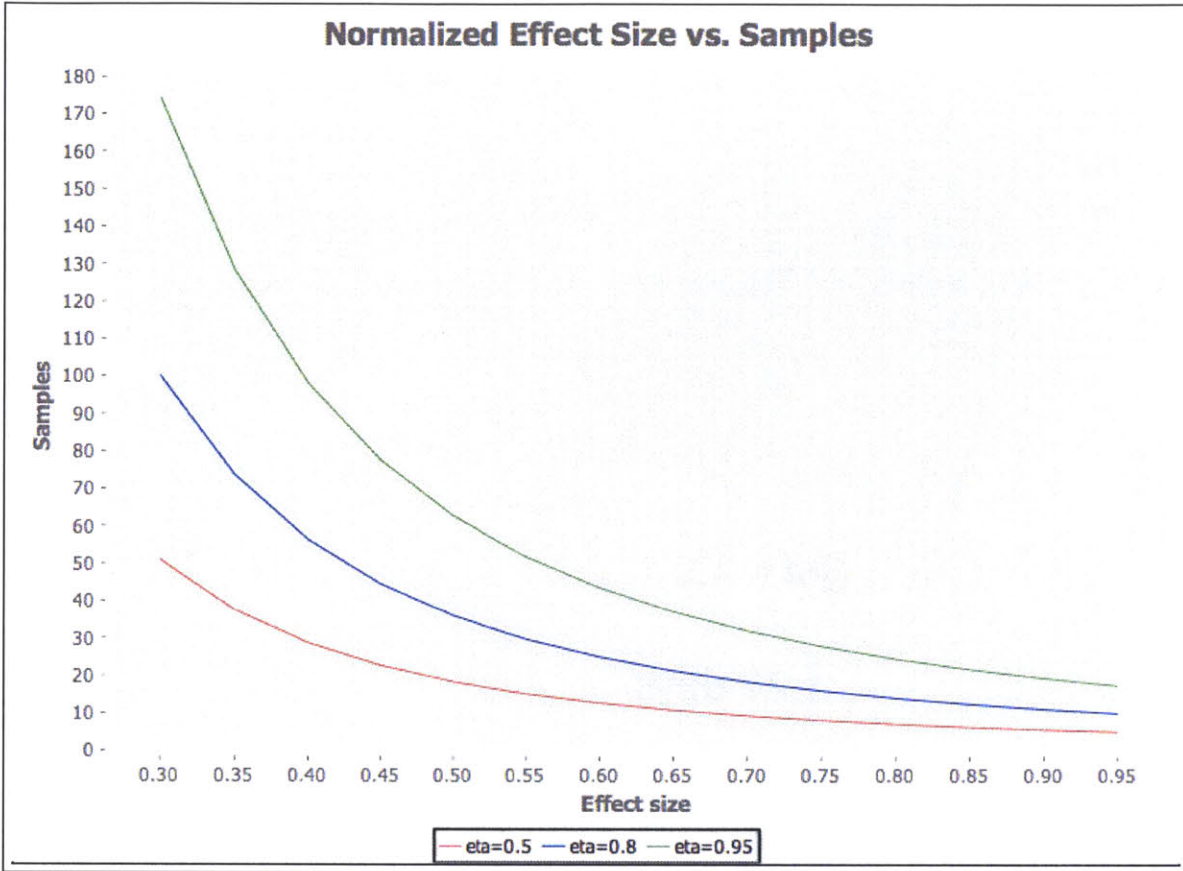


Figure 9-17: Sample size for a two-tailed t-test

of prior information is most significant for effect sizes near 1.0 and below. Large effect sizes can be detected extremely efficiently, as illustrated by the tea tree oil and glycerin examples earlier in this chapter.

### 9.3.3 Improving Designs

One of the most interesting outcomes from the user interviews was that each user started to talk about what really interested them. For people interested in sleep, they really wanted to improve their feeling of “energy.” Many users describe this as dragging by the end of the day and struggling to get through work. Given an average sleep duration of less than 7 hours, this is not surprising!

When we discussed starting new trials, users selected energy and sleep efficiency as the

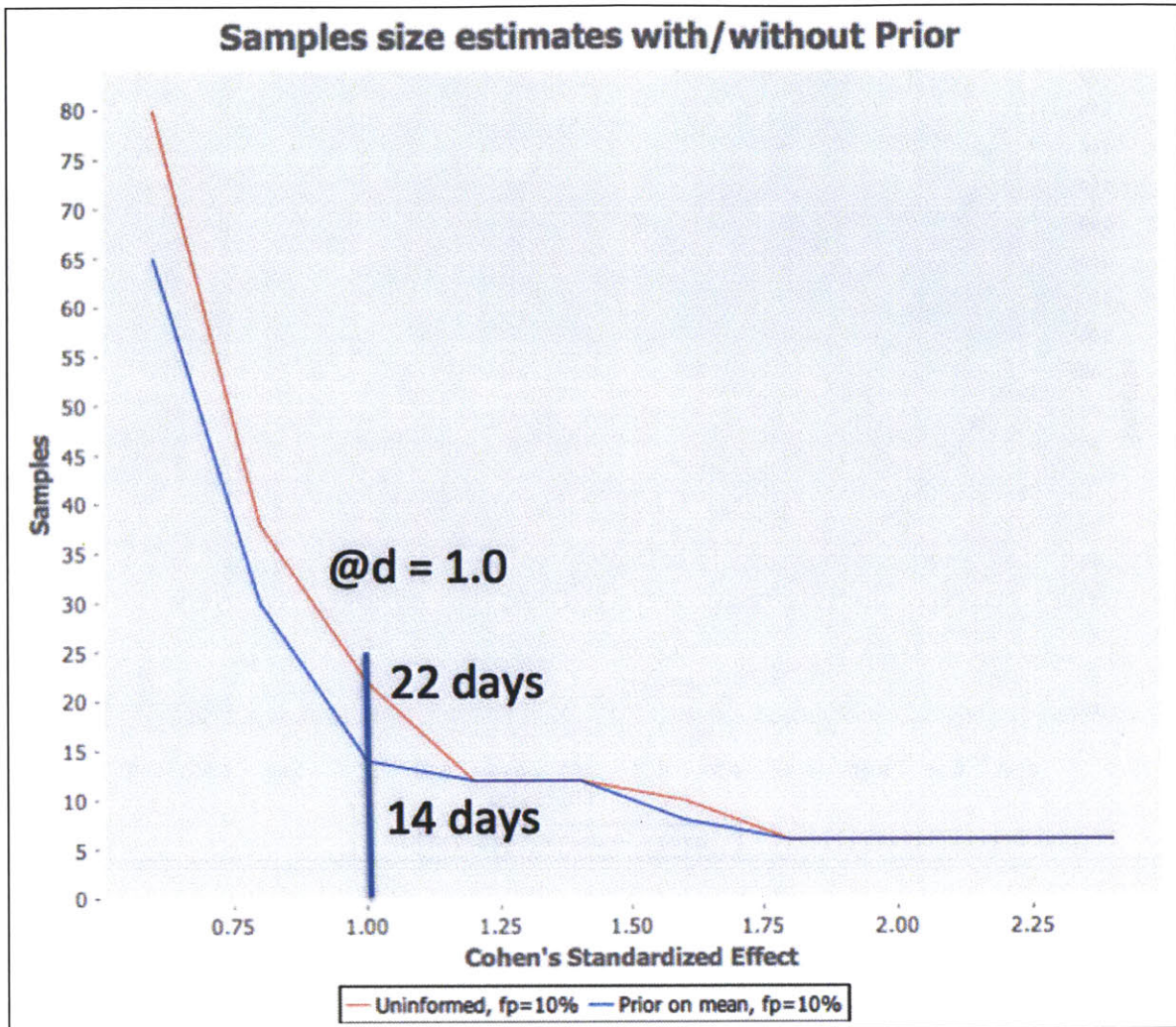


Figure 9-18: Simulated Sample Sizes

more representative outcome measure. This means that they were learning what outcome they truly wanted to improve and focusing on improving that.

As the “Deep Sleep” versus “Deep Sleep Ratio” example above demonstrates, the platform can identify related measures and support users in designing experiments that are more efficient. The use of liver herbs of Section 9.2.4 illustrates how user annotation can identify the rate and magnitude of confounding factors that in future systems can be used to adjust trials to ensure that sufficient power exists to detect the target effect sizes at the desired level of confidence.

## **9.4 Aggregated Self-Experiments will have a Dramatic Impact on Healthcare**

An open-ended question underlying the work reported in this dissertation was whether the data and/or experimental outcomes generated by patients would have an impact on their interaction with the healthcare system.

### **9.4.1 Doctor-patient Relationships**

The first example of the role of personal data in healthcare comes from a mother of a child with Cystic Fibrosis, a serious illness requiring complex day-by-day management and relatively frequent hospital visits. She previously kept track of numerous parameters about her son, but struggled to present it effectively to her physicians. This was largely a bookkeeping problem; she found it difficult to develop the tools to capture and present a set of measures of relevance to her. Personal Experiment's catalog and easy approach to creating SMS-based measures and integrating with 3rd-party services enabled her to do this almost entirely without assistance or guidance from me.

The measures she created and used to track included:

- Airway Clearance - number of vest treatments
- Stool Frequency - number of movements
- Oxygen Saturation - Percentage
- Cough Frequency - A simple 1-7 scale
- Medication Baseline Variation - document the type of antibiotics, if any
- Calories Consumed - Manual estimate of calories
- Appetite - scale of 1-5
- Weight - measured via the Withings Wi-Fi enabled scale and service

A particularly compelling study of the value of this data in the healthcare system is found in a blog post she wrote after she had to take her son into the hospital about a month

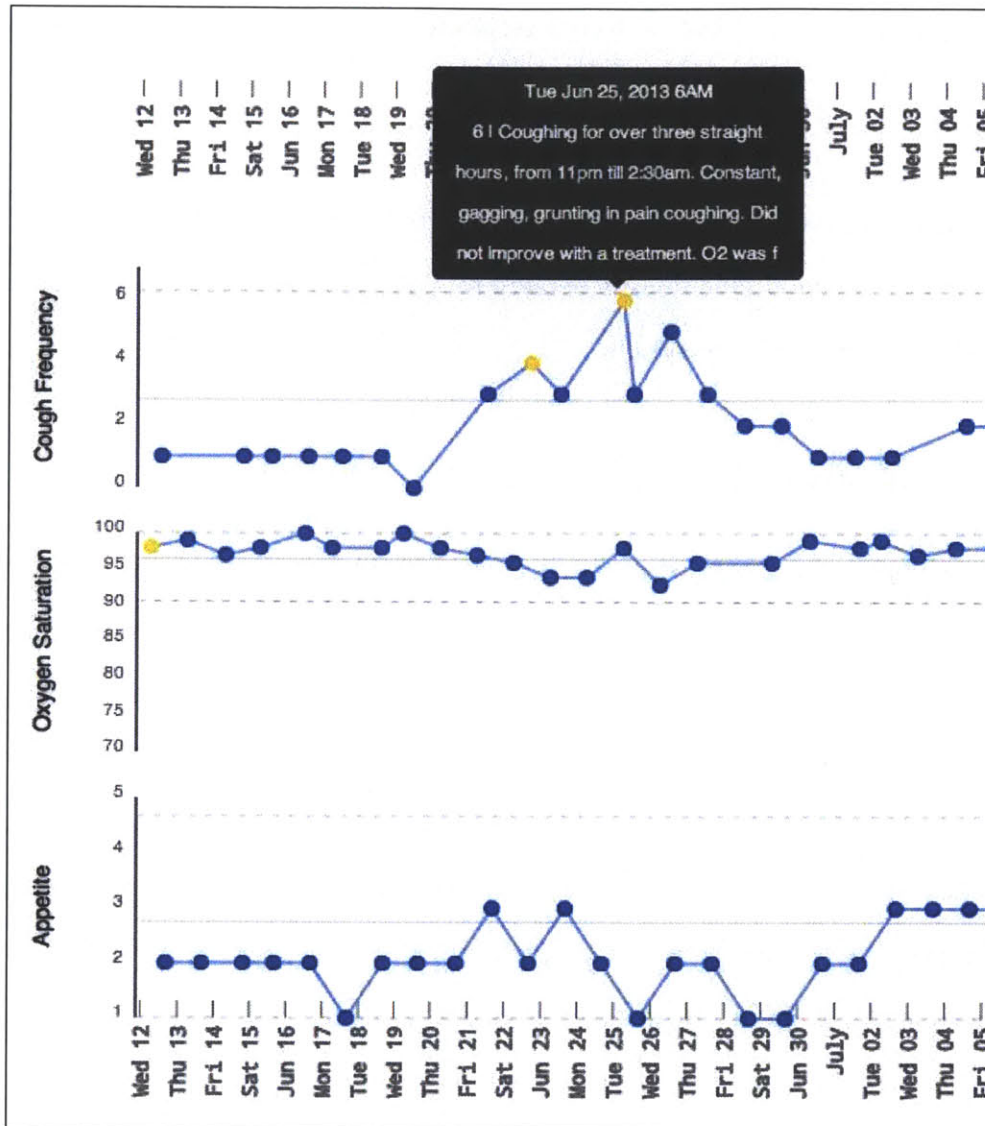


Figure 9-19: Cystic Fibrosis Symptom Chart

after she started tracking his health on the site.<sup>7</sup>

*Once we were up in our room, we were greeted by a resident before we had even been screened by our nurse. She asked to see the data that I had been sharing in the ED (good news spreads quickly!). I shared my data with her and expressed my apprehensions about this being an exacerbation....*

Figure 9-19 shows the data she had available that day. Having data and documenting her experience caused the professionals to treat her feedback with more credibility. Her son

<sup>7</sup>The full text of the story is reproduced in the appendix at Section A.4.

was having symptoms similar to an exacerbation, which is typically caused by infection-related inflammation; however, the pattern this time was different. Because the typical treatment is IV antibiotics requiring a several-week stay in the hospital, she did not want to go down that road unless it was absolutely necessary.

*I did not believe [Son] was having an exacerbation. While I haven't been collecting data for long enough to have caught an exacerbation, I recall him being symptomatic 24/7, and [Son]'s serious symptoms seemed somewhat isolated to sleeping or laying down. I had notes in my tracking that called out the times that these things were happening, and at 3pm he was running around the hospital room and playing without a care in the world - asymptomatic. I thought that [Son] had caught the cold that his siblings had and while he was sleeping/laying, the post nasal drip was throwing him into these coughing fits. His symptoms just didn't fit the bill of his "typical" exacerbation.*

Because it is rare to stop IV antibiotics after starting them, the physicians agreed to observe him overnight before starting antibiotics. She maintained the inhaled steroids she had started him on at home (under the cold, not exacerbation assumption) and shortly after, the test came back positive for Rhinovirus, confirming that the cold was the cause of his symptoms.

*Because I had been tracking his health so closely, because they were able to access his previous test results, because I felt empowered enough to speak up and express my perspective and desired course of action, and because I had given them the evidence to trust me, we found a mutually agreed upon solution that saved us two weeks in the hospital. It saved [Son] the stress of being in the hospital for 2 weeks. It saved me the trouble of trying to arrange my life for a 2 week hospital stay. It saved the doctors and the hospital time and money. My insurance company wasn't being billed for unnecessary tests and an extended hospital stay.*

For more examples of how data and experimentation impact the doctor-patient relationship, see Section 9.4.3 below.

## 9.4.2 Supporting Hypotheses for Research

Several authors have noted that patient communities are an excellent source of hypothesis generation [Swa09] [Rob10]. The challenge for the researcher is separating the unsupported anecdote from a genuine research idea. The Estrogen Study (see Chapter 4) detailed a process by which a community idea was surfaced, turned into a data collection effort, extracted as a dataset, and given to researchers to influence the research agenda. This model could have easily been replicated on PersonalExperiments.org or MyIBD, with the added benefit that theories that involved interventions can also be validated.

Group experiments can provide evidence about a wide variety of factors that lower the risk for researchers by characterizing an observation or pre-validating a hypothesis:

- What is the prevalence of the reported phenomenon?
- How many people document the treatment effect?
- What is the time-course of the response?
- What is the average effect size?
- What are the factors that would influence a clinical trial design?

Providing leads to the research industry in the form of crowdsourced datasets can change the game for advocacy organizations seeking to inject the voice of the patient into the clinical research agenda.

## 9.4.3 The Learning Health System

The Learning Health System represents a vision of the future of medical care, where we systematically learn from every doctor-patient encounter and constantly improve delivery processes by experimenting at the system level using systematic measurement of outcomes to judge improvements. The Aggregated Self-Experiments model personalizes the learning, enabling iterative learning at the level of patients and populations.

The early experience of MyIBD (see Section 8.7) in the ImproveCareNow network establishes the individual benefit and shows promise for population benefits. MyIBD does

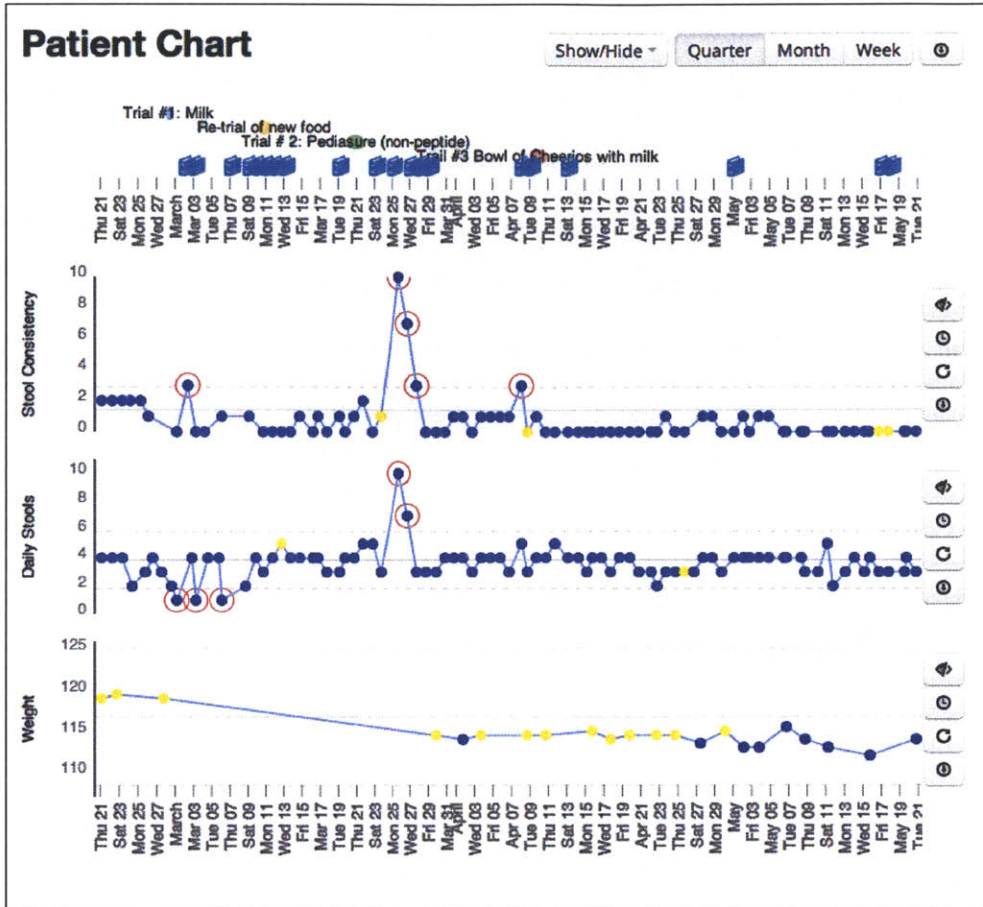


Figure 9-20: MyIBD Ad-Hoc Interventions

not employ the full experimental model described in Personal Experiments. Instead, it focuses on tracking and ad-hoc experimentation to reduce patient symptom variance and identify “special causes” of symptom variation. These interventions are performed by the patient and recorded as explicit intervals during which a treatment was tried. The ImproveCareNow team, at the time of this writing, is evaluating the use of the Personal Experiments experiment model in the provider setting.

### Ad-hoc trials

The patient chart in figure 9-20 shows ad-hoc experiments designed jointly by a clinician and patient. The experiments include temporary introduction of milk and PediaSure into their diet where PediaSure appears to have increased the number of loose patient stools.

The end result was a slow increase in confidence by the patient that his diet would be able to return to normal after a surgical intervention. He was hesitant while the doctor was confident. His first few ad-hoc trials demonstrated a clear ability to resume a normal diet. The general experience of the clinician-patient pairs<sup>8</sup> who use the site in the pediatric care setting indicates that MyIBD enhances the physician-patient dialog, allowing the clinician to focus on asking questions about the documented experience rather than struggling to elicit that experience during the first part of the visit. Both parties indicated generally that more time was spent problem-solving than before the self-tracking data was available.

### **Detailed tracking of outcomes**

Another good example can be found in a patient undergoing an experimental therapy: a fecal transplant. The hypothesis behind this therapy is that disease activity is driven by the behavior of gut microbes, and that replacing the gut microbes can stop the disease process. Figure 9-21 shows the time course of improving symptoms, including the number of times abdominal pain interfered in daily activities (the second time series). The third time series shows a continuous increase in weight, a positive symptom of managed disease. The fourth time series captures the severity of bloating, with lower values indicating decreased microbe activity.

Both the prior two patients reduced the frequency and number of measures, considering their engagement with the site to be a success. They switched into a surveillance mode to see if the gains are maintained over a longer timeframe.

### **Use of measures**

While many researchers focus on formally validated Patient Reported Outcomes (such as the PROMIS [CYR<sup>+</sup>07] measures), these pairings often found value in simple metric

---

<sup>8</sup>The total recruited is over 10 at the time of this writing, heading to a goal of 20. Recruiting was slow due to IRB delays at other sites, a complex consent process, and the difficulty of clinicians adding another step to their workflow. This data will be formally published in 2014.



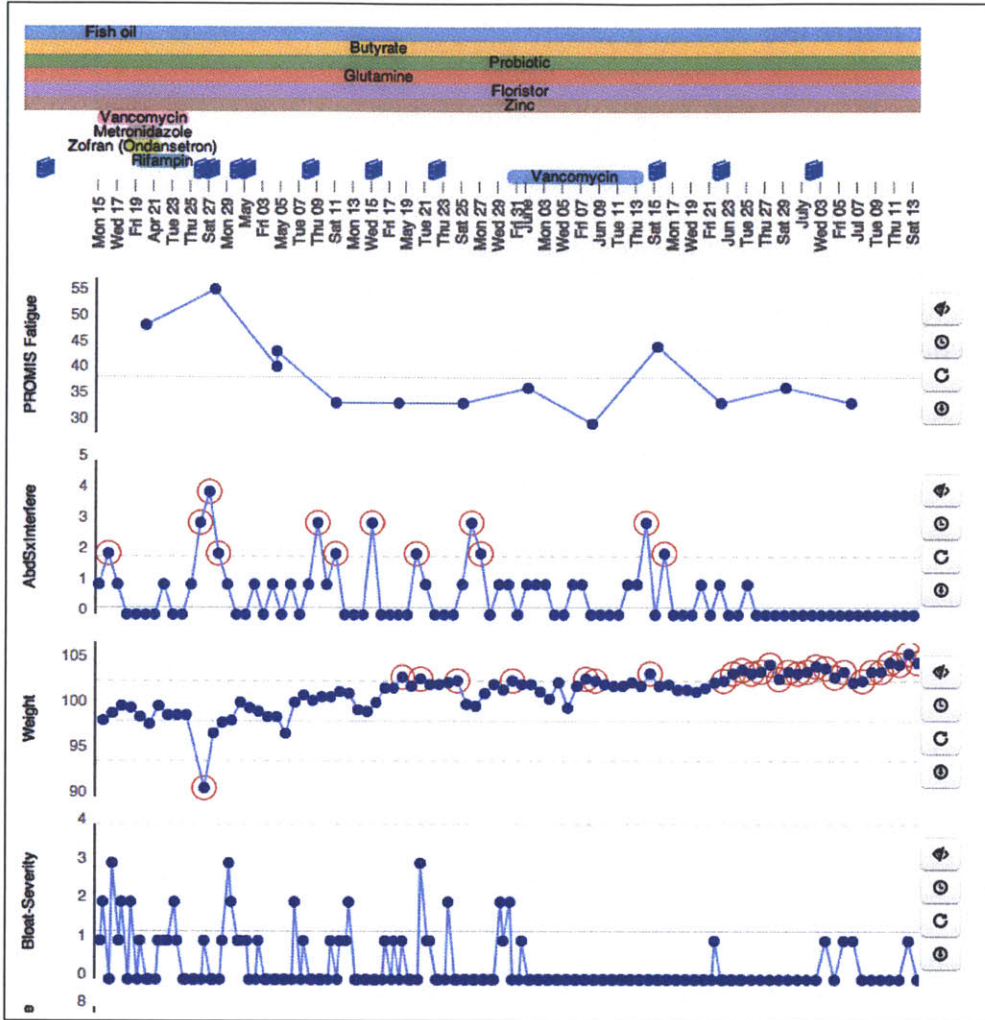


Figure 9-21: MyIBD Transplant Recovery Chart

scales. It is intra-patient consistency, not inter-patient consistency, that is important for characterizing patient variation.

MyIBD did have an interesting case of patient withdrawal. The user, a teenage female, was receiving SMS prompts asking for details about her IBD symptoms. She regularly shares her phone with her friends (to look at pictures, texts, etc.) and a prompt came in that would have been highly embarrassing to her had her friend seen it. It was almost “a disaster” and she asked to stop the study immediately. Prompting via SMS is easy, but it isn’t always appropriate!

## **Prospects for future use**

The concept of Personal Experiments has a great deal of traction with my collaborators in the ImproveCareNow network and elsewhere. MyIBD and Personal Experiments has generated a large amount of real-world interest:

- Cincinnati Children's Hospital and Medical Center has commissioned a second-generation MyIBD platform based on the jointly developed concept of the "Personalized Learning System" that encompasses surveillance, detailed tracking, ad-hoc trials, and formal n-of-1 trials. They also want to explore the group learning potential exhibited by Personal Experiments. This system will be deployed into 10 pediatric specialities at CCHMC and the disease-specific networks in which they take part.
- An adult children's hospital has commissioned a preliminary study of the use of these tools in speciality clinics where they've determined that they can only provide quality care if they know more about what really happens in the patient's life between visits.
- Four funding organizations have explicitly requested funding proposals that incorporate the ideas of MyIBD or Personal Experiments.

The statistical and research methodology communities are beginning to explore the building blocks necessary to apply the Aggregated Self-Experiments framework to the healthcare ecosystem. If healthcare is to realize the vision of the learning healthcare system, upgrading traditional treatment to include self-tracking and/or causal experimentation can improve the fidelity of treatment decisions made by doctors. Aggregated Self-Experiments provides a starting methodology that will address important healthcare questions such as:

- What is the actual time-evolution of the patient experience? (Remove patient recall bias from treatment decisions)
- Does this patient experience a sufficient benefit from the treatment? (Is the effect size what is expected/needed?)

- Who is likely to benefit from this treatment given a diagnosis? (Predictors of heterogeneity)
- Who is likely to have side effects for a given treatment? (Predictors for side-effects)

## 9.5 Observations from Early Prototypes

Exposing users to early versions of the system illustrated that even the use of the terms of experimentation such as outcome, randomization, blinding, etc. is frightening for many users. The technical details overwhelm and discourage them from engagement. Feedback on the first version of the system drove the consumer vs. producer distinction. It also inspired me to reduce the design workflow for consumers to one that is predominantly selection.

After the first redesign, I demonstrated the platform in a live demo session at the 2012 Quantified Self user conference. There were a number of highly motivated self-experimenters there who I engaged in an informal co-design process. The users all articulated the same concern: they didn't want to pay the time it would take to perform a lengthy, formal experiment but were willing to trade off certainty/confidence for "good enough" evidence. This motivated the shift to a more pragmatic, long-term approach to control for and handling false positives from individual trials.

The universal concerns about time burden motivated the emphasis on estimating variance, outcome/washout duration, and relaxing the default confidence constraints while increasing the default effect size target to minimize time while retaining reasonable assumptions. The same feedback influenced the choice of the run-chart control limits described in Section 8.3.2.

One early piece of feedback was that for people facing highly unpleasant chronic conditions, performing a withdrawal phase of a study in the presence of a large effect size in the first treatment phase was unlikely to be something they would do to increase confi-

dence. “If it works, it works.” This motivated support for the switch to the interrupted time series design for treatments with short withdrawal periods where users abandoned the trial to continue the treatment.

# Chapter 10

## Future Work

This dissertation provides a first glimpse onto a rich landscape visible when observing individuals engaging in personal observation and experimentation. The work reported in the prior chapter raised many questions that needed to be set aside to complete a preliminary investigation. For the scientist interested in this work, the infrastructure and instrumentation necessary to support some of the following questions can be quite substantial. Fortunately, a broad array of platforms providing building blocks for personal experimentation are emerging in the commercial and research landscape. Systems like Personal Experiments and MyIBD will also be released as open source projects from which researchers may choose to build.

### 10.1 Improving Trial Design

The central goal of aggregation in Personal Experiments is to improve the reliability and efficiency of an individual experimental trial. The following sections describe additional procedures and algorithmic concepts that may further improve the quality or efficiency of an experiment.

### **10.1.1 Exploiting Covariates**

The current experiment model allows designers to add covariates to an experiment. These are currently only used for manual review of the data to identify special causes or loss of adherence. Future implementations can use the measures to perform multivariate adjustment or automatic identification of better outcome variables (see also 10.3).

To perform automatic multivariate adjustment, the system will need a specification of which measured covariates are independent variables and which are dependent variables for each experiment. For example, in experiments of mood or energy where users are testing treatments such as a supplement or exercise, sleep can be viewed as an independent variable. Of course, sleep hygiene is dependent on mood-related behavior, but it has a direct and measurable effect on mood and energy. To adjust for the effect of sleep on the dependent variables of mood and energy, we need to train a model that evaluates covariation during baseline periods to adjust for negative and positive effects of sleep on mood during treatment periods.

Multi-variate adjustment will often need to account for time-series effects, where the independent variable influences the dependent variable some number of hours or days later. In practice, the use of adjustment is a modeling judgement few users are likely to be able to make, so refinements such as these may require additional support to ensure that assumptions are validated by more expert users.

### **10.1.2 Discovering Carryover Effects**

One of the challenges to experiment design was that few users understood or could make good guesses about carryover effects of a treatment. The current system uses washout periods to drop data associated with carryover from the analysis of a trial. An analytical approach models carryover as a transition effect, for example a linear or quadratic model of the slope of the average treatment effect between the treatment and baseline periods.

A trial design could start with a conservative withdraw period or ignoring withdraw but

padding the duration of baseline periods. Post-analysis of several successful trials could search for the best carryover model that captures the transition dynamics. Future trials would adjust directly for carryover effects and shorten the baseline periods accordingly.

### **10.1.3 Modeling Transitions**

The current approach ignores data collected during the transition between two phases of a trial. Onset intervals are used to define the masking period for transitions from baseline to treatment and washout intervals to mask data collected during treatment to baseline. After observing a number of trials of a given treatment and outcome or covariate pair, it is possible to fit a model to the transition curve such that the information collected during these periods can be used as part of the belief update procedure. For treatments with long onset and washout periods, this may be critical to maintaining user interest. This approach is described with more detail in Chapter 4 of a forthcoming monograph on n-of-1 trials [DEG<sup>+</sup>13].

### **10.1.4 Improved Modeling Fidelity**

The t-distribution used in chapter 7 to estimate the mean parameter value may not be the most accurate way to approximate the true sampling distribution of the data collected in self-experiments. Simplifying assumptions reduce the efficiency with which trials can be performed; noise induced by modeling error must be made up by increasing sample size. Future systems will benefit from more accurate modeling of the distributions of specific measures along with model selection techniques to identify the most appropriate priors to use in optimizing trial execution.

For example, an individual phase's treatment effect may be better modeled by truncated gaussians for scale variables or multinomial distributions for categorical data. These models will influence the minimal sample size calculation, allowing for a more accurate assessment of what the data tells us about treatment effects.

### **10.1.5 Retesting the Baseline**

The analysis here assumes the exchangeability of the user across pairs of baseline-treatment separated in time. In practice, this may not be true. Because most symptoms are multi-causal, a person who is active in their care may improve to the point where the effect size of an earlier treatment is now much smaller. If the patient was better, and is now worse, it may be much larger. Because of this, effect size is in practice a function of the current mean and of hidden variables we may not know about. A safer means by which we may validate exchangeability before performing a belief update step is to ensure that we compare the current and prior baseline profiles. If the models are similar (i.e. could perform a null hypothesis test between the two baselines), then we can assume exchangeability and simply update the belief model with the latest observations of treatment and baseline.

If the baselines are different, we may want to change the prior belief about effect size. If the users are doing better, then their likely effect size will be smaller, requiring more samples to detect an improvement.

### **10.1.6 Empirically Chosen Effect Size Categories**

Cohen's breakdown of standardized effect sizes for clinical studies is not universally valid, and may be too aggressive for the Personal Experiments settings where larger effect sizes are targeted. By sampling the opinions of trial participants, observed effect size could be mapped to user assessment of effect importance (e.g. small, medium, large) that is likely to match up with the next patient's subjective opinion about the relevance of an effect.

## **10.2 Alternative Trial Designs**

A second area for future work on individual trials is extending the architecture and data models described in the thesis to accommodate more creative trial designs that will save user time and energy.



### **10.2.1 Short Onset/Offset**

The current architecture is designed around batch studies, sequences of days separated by several days of onset and washout. However, many treatments, such as Seth Robert's one-legged standing [Rob10], have immediate effect, and no known long-term effects. Thus, a trial design that alternates day-on, day-off would be more effective because longer-acting confounding factors would act equally on each pair of measures, significantly improving the value of randomization for these kinds of treatments and reducing the total amount of time needed for a trial.

Similar designs can accommodate psychological trials, such as the effect of caffeine on mental performance. Pairs of trials, one before coffee and one after, can be done within the same day over several days to similar effect.

### **10.2.2 The “Two Armed” Body**

Topical treatments that address local skin conditions, such as the glycerin and witch hazel treatment described in Chapter 5, can use two different parts of the body for a treatment arm and the control arm of a trial (or the two arms of a subject!), rather than measuring the body over time. This dramatically shortens the trial, removing onset/washout requirements and all baseline conditions. It also perfectly smoothes over confounding factors so long as no other topical treatments are used, or are used equally.

The challenge for this design is changing the tracker architecture to allow two trackers, each customized to the variable in the control arm or the treatment arm.

### **10.2.3 Factorial Design**

A factorial design is used to assess the effect of a set of treatments alone or in combination. It takes a set  $N$  of interventions, and groups them into  $M$  overlapping subsets. Each phase of a trial consists of a test of one of the  $M$  combinations. The post-trial analysis looks at

the association between each therapy and the magnitude of the outcome response to assess which intervention or combination of interventions had the greatest effect.

The benefit of factorial designs over the sequential model proposed in the thesis is the ability to detect synergistic interactions between multiple treatments. These trials are much harder to develop a standard workflow for, harder to adhere to, and difficult to present visually.

#### **10.2.4 Dose-Response Trials**

One user designed a dose-response trial using a medication period and two experienced side effects and reduced the dosage of a supplement. Introduction of a dose-response trial model would be valuable for users who want to find the best property of a treatment. Sleep is another good example of this. Many people want to improve their sleep hygiene, but can't figure out the optimal amount. If the impact of sleep on an outcome measure is non-linear (e.g. fatigue, mental performance or mood), then a dose-response trial could help the user identify the optimal amount of sleep.

Dose-response trials could be done in two ways depending on the nature of the treatment and the measure. For sleep, given that the hangover effect of a short night of sleep is short-lived, a simple correlation of the outcome measure and the sleep values with an auto-correlation adjustment could be used to fit a curve and identify the point of maximum gain. For something like melatonin, a series of short periods at different doses should be sufficient to identify a minimal effective dose or possibly the parameters of dose-response model.

#### **10.2.5 Dynamic Trials**

As mentioned in Section 9.1, users are better at reporting observations and recognizing patterns in data than creating solutions. To enable more users to act as producers, we can reduce the burden of trial design by allowing trials to be fully dynamic processes through

which the system probes the state of the user under different exposures. The frequency of alternations, and the duration of the trial, is driven by the model of Chapter 7, but automatically learned from prior trials.

## **10.3 Surrogate Measures**

In the current system, each measurement object for a given variable is predictive of the others. Whether total sleep is measured by a device or by memory affects the reliability, variability, or likely adherence of the measure, but all measures will have high correlation to one another.

As Patient Reported Outcome (PRO) measures become more central to medical care, and systems like Personal Experiments, PatientsLikeMe, and medical device applications encourage users to be more rigorous in their observations, there will be opportunities to identify the standard relationships among variables. Proxy variables reliably predict the response of another variable to a class of interventions, but have different properties such as a faster response time to treatment, and ease of measurement.

For example, in the case of our hypothetical patient Alice, her fatigue was correlated to her psoriasis activity but it responded within days, rather than weeks, to a dietary intervention. If the Pagano diet improves fatigue, then the Pagano diet over the longer term is likely to impact more slowly varying symptoms of psoriasis. The system could represent these associations as predictive factors from one point in the hypothesis space to another, thus preferring experiments with faster responding outcome variables that predict the response of the primary variable of concern to the user.

### **10.3.1 Surrogate Measures for Clinical Outcomes**

One challenge of running trials without an intimate connection to a clinical care delivery infrastructure is the lack of laboratory testing. Laboratory tests are often more reliable than

user-reported measures. Self-experimentation would be greatly enhanced by identifying more accessible and practical measures that are predictive of laboratory measures. User communities could discover these measures by comparing their laboratory data, building a population dataset that assesses how well one measure predicts the other.

### **10.3.2 Passive Measures**

Passive measures are the easiest for users to adhere to. The work of Madan and others demonstrated that location and accelerometry data are predictive of mood, pain, and other variables of interest to chronic disease management [MCM<sup>+</sup>12, RACB11]. While this work is primarily being performed using device-based data, would even less intrusive measures of online activity, such as Facebook interactions or e-mail sent to friends, also predict mood or pain?

Identifying simpler, passive ways of measuring important variables will broaden the accessibility of personal experiments. In the ideal case, an experiment simply becomes taking treatment when reminded for a short period of time, removing data collection burden from the equation.

## **10.4 Combination Treatments**

Many therapies explored by users in practice are complex combinations of multiple treatments. If a given user responds to the compound treatment, what is the component of that treatment that is responsible for the observed effect? The aforementioned factorial experiment performed by a single individual can certainly help, but at a heavy time cost for an uncertain gain.

Instead, a group of like-responders could be allowed to volunteer to perform a distributed factorial experiment: to try one sub-combination of the therapy to see if they still respond. Performing cross-overs between arms based on the results, similar to the far more

sophisticated approaches of modern clinical trials such as the I-Spy 2 trial [BSK<sup>+</sup>], might yield an efficient procedure for small amounts of individual volunteer effort to discover the causal agent in combination therapies.

## **10.5 Accommodating Global Constraints**

The theoretical model presented by chapters 6 and 7 frames the n-of-1 trial as a decision problem. Markov Decision Processes [KF09], including the computationally explosive Partially Observable Markov Decision Process (POMDP), were developed to represent sequential dependencies between decisions to order decision problems under constraints and uncertainty. Given user preference constraints, future research should investigate several interesting phenomena:

- Performing experiments to increase the knowledge for a future decision, to reduce the overall time to treatment,
- Impact of policies choosing between learning about new treatments and maximizing patient value from known treatments (multi-armed bandit problem [ACBF02]),
- Allowing users to volunteer to try new treatments to improve learning, and
- Identifying when the probability of future reward is unlikely.

## **10.6 User Interface Elements**

### **10.6.1 Tagging Taxonomy**

Collaborative development of tagging taxonomies in health has been reported in prior work [SW08]. Although collaborative tagging of catalog elements in Personal Experiments was limited, tracking of search terms and click behavior could facilitate the automated discovery of useful tags. Users could be asked occasionally to approve or reject new candidate tags. Adding suggested tags to an auto-complete drop-down to give the users suggestions

on likely tags makes this approval implicit. An improved and more complete tagging taxonomy improves searching, filtering, and recommendation functions as the current framework implementation uses tags in place of conditions (what an experiment is good for) and custom search terms.

### **10.6.2 Medical Taxonomy**

Similarly, were formal medical taxonomy to be utilized, developing the mapping of suggestions between informal tags and formal elements of a taxonomy could be performed by sampling users with questions like such as “Is the measure X the same as taxonomy element Y?”

### **10.6.3 Variable Control Limits for Run Charts**

The heuristics necessary to compute the current baseline on an ongoing basis are very specific to a measure. For example, how do we separate a new baseline from a temporary special cause influence? Knowledge about the typical long-run behavior of a measure is necessary to determine the appropriate window size for adjusting limits. In the future, these factors may be computable empirically from population data.

## **10.7 Integration with Health Social Media Forums**

One proposal discussed with several forums, but not implemented at the time of this writing, was finding ways to embed aggregated experimental outcomes into existing forums as a way to augment the validity of assertions made by users. This can be done by tagging existing topics or even specific cause-and-effect assertions about treatment or other intervention efficacy. Embedding serves as a recruiting tool, bringing people from existing forums into Personal Experiments to try something out for themselves, rather than trying to build communities from scratch.

There are several ways in which contextual information from a site like Personal Experiments may be presented. A summary widget next to the main content, a popover, or a hyperlink annotation on treatment expressions such as those identified by the techniques of Chapter 3.

Embedding can be automated by adding a small piece of javascript that adds links, or populates page widgets based on an analysis performed of the content by a crawl of the site. Such techniques are commonly used by advertising systems and may have similar discoverability benefits for patients seeking more information. Links can flow both ways; forum posts that are annotated can be linked back from the inside of Personal Experiments.





# Chapter 11

## Conclusion

Millions of people exchange treatment anecdotes online every year. Aggregated Self-Experiments empowers an individual to structure their experiences so they can apply the tools and processes of science to evaluate these unstructured anecdotes in support of better self-care decisions.

PersonalExperiments.org is the first platform to implement a personalized model of experimentation that enables individuals to act as scientists. User response to the concept of experimentation was strong with most users learning from their experiments and 95% of study participants expressing a desire to continue experimenting.

The framework supports explicit documentation of an experiment. This enables replication of the experiment. Repeated trials of an experiment provide many benefits. I show that:

- Users can replicate an experiment with their own personal trial,
- Replications within a person increases their confidence that a treatment causes a change in outcome,
- Replications across people train models that make it possible to run faster trials,
- Population outcomes can be used to prioritize among multiple experiments for the same outcome or condition, and

- Engaging with an experiment educates users about the process of experimentation.

More importantly, the framework generates data that is useful for healthcare professionals and consistent with trends to develop a Learning Healthcare System that learns from interactions that happen in the course of ordinary care, and now, daily life. My work with healthcare providers has demonstrated significant conceptual impact and practical adoption. Specifically, I've shown that:

- Patients can use data to present their case and argument to practitioners,
- Patients and caregivers can design and engage in ad-hoc trials of therapy using their everyday experience,
- Biomedical researchers can be motivated by patient-reported population data to consider new hypotheses, and
- Researchers, and increasingly provider organizations, want to explore rigorous models for learning from data collection from patients between office visits.

For developers of patient-facing self-tracking tools, I've identified a set of design principles which should govern the development of tools that engage users through lightweight experimentation:

- Decomposition of experiments into a space of treatment-variable-design,
- Strategies to prefer characterization and long-term adaptation to the conventional approach of rigorous up-front control,
- Algorithms for computing the parameters of practical trials, and
- Explicit feedback loops that improve experiments and parameter estimation over time.

I've also identified a broad set of topics for future research in methodology, statistical modeling, the design of real-world measures, and user interfaces.

This dissertation asserts that Aggregated Self-Experiments, and other approaches built on the concept of lightweight, end-user experimentation, will help self-tracking and social health companies make their data more actionable. Experimental outcomes collected at scale will be a key enabler for the future of personal and biomedical discovery and learning

from patient experience at the “point of care” will accelerate the transition to a Learning Health System.



# Appendix A

## Social Media Content

### A.1 Edison Experiments

<b>ID</b>	326
<b>Title</b>	Getting rid of shin slints
<b>Description</b>	I will keep a diary to track pain, treatments and prevention exercises.
<b>Completion Test</b>	It will be a success if I have no pain on 14 consecutive days.
<b>Journey</b>	I will enjoy running much more when I am injury-free.

<b>ID</b>	246
<b>Status</b>	Running
<b>Title</b>	Strength Training
<b>Description</b>	Use one strength training method on Left and one on the Right. Baseline measurement of L and R biceps at point 3above elbow crease, mark dot @ 3with perm. marker; update as needed. Record on calendar baseline weight. Left arm: Start with heaviest weight I can do 3-4 slow (10 sec. each motion) bicep curl. R arm: 12 reps with conventional 4 count each motion. In addition to any arm work in my cardiodance classes, I will do these arm exercises every 2 to 3 days (3x/week), adding weight in small increments on the side on which it becomes easy
<b>Completion Test</b>	When I regain enough strength to do each side at 20 pounds using the assigned method. Then, if I am motivated and ambitious, I will decide if a switch is merited. I may do another experiment on maintenance.
<b>Journey</b>	I am doing something good for myself that costs nothing, is fun, and it just might settle the issue.

<b>ID</b>	228
<b>Status</b>	Completed
<b>Title</b>	<p>Effects of drinking milk on night sweats” ”Description: I worked out a while back that if I drink a lot of milk just before bed I have night sweats after 5 or 6 hours of sleep. A fair bit of tracking went into working this out!</p> <p>I’m interested to know:</p> <ol style="list-style-type: none"> <li>1. Where the milk threshold lies in ml before I start to experience sweats</li> <li>2. The effect of the amount drunk on the severity of the sweats</li> <li>3. Does it make a difference whether I drink the milk alone or with something else (i.e. breakfast cereal)</li> <li>4. The time to onset from drinking to night sweats</li> <li>5. Whether there is the same effect with unhomogenised milk as there is with homogenised milk.</li> </ol>
<b>Completion Test</b>	I will have answered my 5 questions.
<b>Journey</b>	Just the idea of this experiment makes me laugh. I can’t wait to see my wife’s face when I tell her what I’m doing. Not happy!

<b>ID</b>	47
<b>tatus</b>	Runnin
<b>itle</b>	<p>Increase compliance with dieting program with one (maybe) simple behavioural change  Description: Since my daughter was born nearly 3 years ago, I have been trying without success different formal and informal dieting programs to lose the weight I gained during and after pregnancy. Recently I've been reading (via GTD Connect) about a new book by Roy Baumeister on Willpower, which argues that willpower is both a finite resource and like a muscle that needs to be strengthened (haven't read the book yet). It occurs to me that my lack of dieting success is partly due to my attempts to apply willpower in order to change too many things all at once (follow a food plan, drink more water, take vitamins, drink more water etc etc). Some of these components are behaviourally quite complex and, I think, need my sole and sustained focus to develop new habits; I've heard that new habits take at least 21 days to develop, if not longer. I have read also that willpower is used up or weaker (depending on your metaphor above) when your life is challenging, you don't get enough sleep, or eat poorly. These are all constant conditions of my life as a working mother. So my general hypothesis is that I will have more willpower if I focus on just one behavioural change at a time amidst my very busy and complex life. Since willpower is impossible to test directly, I'm operationalising it in terms of my success on a weight loss program following one specific behavioural change, which I intend to implement for at least 21 days. My prediction is that if I engage in this specific behaviour for at least 21 days, I will form a new habit (which may start to become automatic) and my weight loss success will increase.</p>
<b>Completion Test</b>	<p>I will track my compliance and weight loss over at least the next 21 days (the number of days that some people say are required to form a habit). I will know that I'm done when preparing my food and following the diet program become less of a struggle (more automatic behaviour) and I slip off the program less frequently (when the average number of days compliant per week is high). I will also know I'm done when I feel as though I'm ready to move on to the next behaviour I want to change (last week it was stopping drinking diet coke, but I didn't know about Edison then!).</p>
<b>Journey</b>	<p>I will enjoy increased feelings of control and success. I will also enjoy noticing if my ideas about willpower are confirmed by my results.</p>

<b>ID</b>	476
<b>Status</b>	Completed
<b>Title</b>	Breakfast improvement” ”Description: My normal breakfast: Oatmeal, yoghurt, raisins, almonds and an apple. My sample meal was 688,74 calories. For the next four days: Chicken, beans, tomato and onion. 479,1885 calories. The experiment consists of seeing if I can consume less energy and stay full for longer. Normally I can go from 7 to 10 before hunger kicks in - thats 3 hours.
<b>Completion Test</b>	After 4 days
<b>Journey</b>	If I find an ideal meal I will be happy.

<b>ID</b>	503
<b>Status</b>	Running
<b>Title</b>	When to take One a Day Vitamin? In the morning?” ”Description: I want to begin taking One a Day vitamins; given my busy schedule, eating lunch (or breakfast...or dinner) is sometimes sporadic and not necessarily the most healthy. As such, supplementing with a daily vitamin makes sense.
<b>Completion Test</b>	After 30 days in a row of successfully taking a vitamin (or until it becomes a continuous habit...)
<b>Journey</b>	By using Edison and keeping track of my progress! Love it!

<b>ID</b>	204
<b>Status</b>	Completed
<b>Title</b>	Memory Loss” ”Description: Remove Dental Amalgams
<b>Completion Test</b>	After all the amalgams are gone, (expensive... have to do it over a few months) I will wait a while to see if my memory improves. or measure it as I go along on the journey.
<b>Journey</b>	Having dental works is not enjoyable. Remembering better is very enjoyable.

## A.2 Patient Experiments: TalkPsoriasis

### A.2.1 Slippery Elm Treatment

Posted on July 25, 2011 to TalkPsoriasis.org

In my search for a livable solution to my very severe Psoriasis I have begun searching for ”Cause and Effect” instead of the ever more popular ”Symptom Treatment” that I have personally participated in with disastrous results for the past 45 years.

I have reread all the theories and my analytical mind has assisted me in selecting two causes for this ailment. I believe fully in the leaky gut syndrome



however I must also give credence to chronic inflammation as my body seems to exist in this state. The treatments in my case should overlap I think.

I began this trial July 19th 2011. and have been updating my condition once a week in photos.

I am doing the following...

5 Slippery Elm caps emptied into water, allowed to sit 5 mins then consumed.

Blue berries mixed with 2% cottage cheese every morning and flax seed oil directly behind it.

Gluten limited not fully free but mostly.

Limiting fats except those in flax, fish, and primrose oil.

Taking a probiotic supliment

Soaking in salt water 1-2 times a week time permitting later in the day so I dont burn. Left salt on my body 12 hours.

Avoiding processed foods like the plague.

I awoke this morning with a major over night healing. I guess I neglected to say that I am currently 3/4 covered. The P has attacked my fingers and nails for the first time. Both thumbs, index on left hand and pinky on right. This morning my body is mostly smooth and flake free. My P is white with small spots of red scattered but limited.

Something is working. I have been photographing my arm and hand which is covered. I will photograph again today even though I'm not due to again until Thursday. I'm happy to share these photos but Im new here and still unsure how to work things on this site yet. I hope this info may help someone else.

### **A.3 Patient Treatment Advice: TalkPsoriasis**

This transcript was copied from the TalkPsoriasis forum on March 18th, 2012. Usernames, URLs and names were anonymized.

**Alice** – *Subject: Treatments for other symptoms like fatigue, brain fog and lethargy*

*I'm wondering if anyone else has symptoms that go along with their psoriasis similar to my own. I have the most common form, stable and mostly on my torso/scalp.*

*Over the past 10 years I've found that these symptoms are getting worse. I often wake up feeling like I got no sleep, am sluggish, with puffy eyes, sometimes blurry vision that lasts for hours and horrible lethargy. When these symptoms are bad, I also have trouble remembering things, speaking clearly, and feeling emotionless (not really depressed, just without feeling). The funny thing is that even when I'm feeling like this, I can still go be athletic (running, cycling) although sometimes find it hard to complete a workout on my worst days.*

*Two years ago I tried an elimination diet to remove dairy, gluten and all sugar from my diet (except allowing some starches). I was almost paleo, with limited veggies, lean*

protein and fats making up most of my diet. After 3 weeks of that I felt really incredible and since then I've been trying to figure out a diet that is easier to maintain over the long term without much success. If I have alcohol, lots of sugar, etc - then I 2-3 days later I'll have some horrible days and my P will act up (more itchy, scaly, etc).

My base diet now is gluten and dairy free, but I allow starch, fruit and modest amounts of sugar. It hasn't been working terribly well. I believe that I've seen some benefit from B-vitamins, tumeric, mlik thistle, chicken broth, and magnesium/calcium supplementation, but I don't know for sure.

Every so often my psoriasis will start to fade for awhile while I'm trying to restrict my diet, but it's hard to sustain! I'm considering going back on a full elimination diet for awhile.

Anyone else have these symptoms? What have you tried? How has it worked?

**User1** – You may be low in vitamin B that helps us with a lot of those symptoms. I take a B supplement twice a day and helps me. Good luck.

**User2** – Hi! It seems that you are receptive to homeopathic/natural remedies, so I thought I'd reply. I have just started seeing a homeopathic physician. I had a battery of tests: blood, hair, saliva, urine, etc. I am an otherwise healthy individual, but this is what the tests showed and the supplements he started me on:

low thyroid and adrenal function I don't make enough HCL in my stomach and have issues digesting food calcium and magnesium deficient potassium deficient high silver toxicity

Vitamins/supplement:

natural vitamin/mineral capsules calcium/magnesium capsules 2 thyroid meds 1 adrenal med trace mineral drops super enzymes omega 3 fish oil

Since I just started this regimen (along with the Pagano diet), I have more energy and feel closer to "normal" once again. I used to be so tired that I took a nap whenever I had a chance. I also had the emotional "numbness" that you speak of. It is getting better. I just can't believe that just a few supplements can make such a difference. I always hesitated because a pill is a pill and I hate taking them. But my homeopathic doc is in his late 80's and said all the folks that used to pick on him (call it a sham) for taking so many supplements are long dead-he is the last man standing that takes no prescription pills and has no medical issues today...

If you have the economic means and are interested, perhaps visiting a homeopath would help with your symptoms...

I wish you well!

**User3** – Blurred vision not a good symptom. Go see doc! Could be a lot of things or you starving yourself ( low sugar) take care of yourself! there's a herbal in vitamin section of Walmart called LIVERITE , fab for brain fog!!!

**User4** – I used to feel that way all the time and when I eat a lot of gluten or sugar, I experience it again. what has helped me the most is juicing. I make a fresh mostly green vegetable juice every other day. ON the off day I make a green monster smoothie with Kale. the enegy you get from these is amazing.

**User5** – Yes I have those symptoms too! Today seems really bad I ate pizza 2 days ago and it is kicking my butt today. Diet has a big effect on me. Like you there are days when I am almost to the point of narcolepsy if I am not active. B vitamins seem to help a lot. I

*am about to try an elimination diet again also. User2 your regimen looks good I need to find a homeopathic/naturalpathic doc. The medical doctors have cost me a fortune and no results sad to say.*

**User6** – *Hello! I have very similar symptoms and have tried very similar things. I also feel amazing on a gluten and dairy free diet. What I learned on the allergy elimination diet was that I was mildly allergic to eggs and corn. Are you eating corn? I find that it makes me very lethargic. I have started acupuncture which has helped my sleep issues and my skin. My acupuncturist specializes in skin issues. She recommended a chinese herb blend for my insomnia and I highly recommend it. It doesn't make you drowsy, but stops the "active brain" which is what keeps me awake. The blend is also good for your skin and has kept my legs clear. It is called "Dr. Shen's Good sleep and worry free" - google it and you can buy it online. It is safe long term as some Chinese herbs are not - it does take a week to get into your system and start working. I'm not sure your age, but wondering if some of my symptoms are pre- menopausal - ugh!*

**Alice** – *Thank you all for the comments and suggestions.*

*@User1 and @User3: Oddly, I react badly to most multi-vitamins and comprehensive B vitamins. It's really weird, but it makes me feel quite a bit worse in the hours following it. I can and do take B12 and B6 in isolation and that definitely seems to take the edge off it, but not as much as that restricted diet. Recently I tried to add B1 and Biotin, suspecting some issue with nutrition that's affecting the Krebs cycle, but I'm not sure yet. If B vitamins do help me, especially all the veggies I eat, then that suggests I might have some problems with proper digestion, rather than just a leaky gut kind of problem (i.e. bad stuff gets through but good stuff doesn't?)*

*@User6, I do eat corn and that might be an issue. I feel really good eating stew which has a little bit of corn, but polenta seems to hit me hard. I've re-started a complete elimination diet this week so can do some allergy tests. I'll try removing eggs and corn for the cleansing phase, although I do love eggs! FYI - I'm nearing 40 but am not likely to experience pre-menopausal symptoms due to my gender. ;)*

*I'll reply to some of the other suggestions later!*

**User6** – *Everyone should be reactive to corn They splice the seed it with glyphosphate( roundup) . then we all eat the weed killer. taking the farms away from farmers has enabled us to become victims of mass production food poisons. ! Hey I eat the crap too, cannot afford to go to all organic ( half of that's a lie too) But in the future there will be many many more food/ health disorders . Sad legacy we leave!*

**User7** – *I have the same fatigue, trouble speaking and finding words. My vision gets blurry too with a bad flare. I'm allergic to eggs, react to corn, gluten, and dairy. I also have trouble with food dyes. Especially yellow dye. I have found that if I follow my diet 90% of time, my body seems to handle the 10% of bad foods I eat. I pick and choose wisely when I know I will probably cheat on the diet. I also had to remove all citrus, tomatoes, red meat, and alcohol from my diet. I know it sucks!! However, being able to function somewhat normal, is worth the sacrifices. Don't get me wrong, there are days I just don't care. Then I have to pay the price a few days later for eating the wrong things!! Best of luck. Stay strong. I take sulfasalazine, vitamin d, b vitamins, zinc, selenium, vitamin c. I can't take a multi vitamin either.*

**Alice** – *@User7 You sound just like me!*

*Well I'm on day 3 of eliminating most sugar / carbs from my diet and I started taking B-vitamins at night. I'm been slowly feeling better and less like I've been hit by a truck. Fewer cravings on the diet this time around, but boy is it hard to adjust to snacking on nuts and veggies.*

*An example day: my breakfast was basically "bulletproof coffee" plus some almond butter. My lunch was rotisserie chicken, poached/steamed kale, mixed veggies, some mashed red potatoes, and a touch of fruit. Almond butter for a snack. For dinner I had a shake of fruit, veggies, and protein powder. Limited sugar except fruit, minimal starch, etc.*

*FYI - I had some supplements available that had many of the components of the recommended Liverite product (phosphatidylcholine, NAG, inositol, etc) and I'm definitely much feeling better this afternoon after taking them, though too soon to say if it's random, caused by the dietary change, or the supplementation (or a combo!)*

**User5** – *Question, are any of you taking milk thistle or dandelion. These herbs really help me. They help keep toxins out of the blood. If I slack on taking these I have problems. My health practitioner works with pregnant women and they will have liver problems that cause break outs with skin problems similar to P. They can't take medication so herbs and diet are all they can do.*

**User7** – *Blurred vision, foggy brain, sluggish, lethargy. Yes I thought that was premenopausal too. I have since discovered that I only feel like that after eating things like junk food, processed foods, dairy, gluten and sugar. I have noticed that since I have started to follow the Dr X nutritarian way of eating I no longer have blurred vision, foggy brain, etc...For example, Dr X urges that 50% of your vegetables be eaten raw and the other 50% cooked. Prior to starting Dr X's way of eating I never ate any of my vegetables raw. And in fact ate very few vegetables at all. Less brain fog and clearer skin are just two of the benefits. Here are some links: [URL1] [URL2]*

## **A.4 Blog post by a Personal Experiments User**

**Reproduced with Permission, original posted July 2, 2013**

*A few months ago, I shared a story with the [Foundation] about my vision for the future of chronic care management. If you missed it, you can read it here. Last week, I lived my "what if". Let me explain.*

*About two weeks ago, we packed up shop and headed east for our summer visit to the grandparents house just outside of Philadelphia. About a week into the trip, [son] developed a cough. He'd been doing pretty well and so this cough was new but not anything that I was terribly worried about. Two of his siblings had runny noses and coughs and I'd figured he had just caught the virus that they had. After a few days, the cough frequency was on the rise and his oxygen saturations were on the decline. I track many of his symptoms/behaviors with a tool called PersonalExperiments so I could see from his data that this was the case. I put in a call to his Pulmonologist back in [town] suggesting we start an oral antibiotic and after some discussion she agreed that that might be the best next step.*

*Later that night, [Son] went into some serious coughing fits like I had never heard before. He would cough and gag for hours on end, getting little to no relief from breathing treatments or airway clearance. He finally fell asleep around 2am, but woke up worn our*

*and not much better. The next day he had another one of these intense coughing spells, and his doctor and I felt that he might benefit from an oral steroid as his symptoms seemed to indicate that there may be some serious inflammation causing the intense coughing fits. That evening, we started both the steroid and oral antibiotic and he went to bed. Then, at 11pm, 2am, 4:30am and 6:30am he was up again, coughing and gagging, still getting little relief from breathing treatments, so we decided that it was time to head to the ER.*

*We arrived at [Hospital] around 11:30am and were immediately taken back into a room. The doctors and nurses listened to my whole story and I shared with them the data I'd been collecting so they were quickly able to see what had been going on. I also gave them access to his EHR through MyChart, where they saw his bacteria sensitivities from his most recent culture and decided which antibiotics would be most effective in treating his exacerbation. We had checked in, gotten an x-ray, done a viral panel and an epiglottic culture and were admitted and moved up to a room in under 2 hours. If you've never had the opportunity to visit the ER, that quick of a conversion from arrival to admission or discharge is unheard of.*

*Once we were up in our room, we were greeted by a resident before we had even been screened by our nurse. She asked to see the data that I had been sharing in the ED (good news spreads quickly!). I shared my data with her and expressed my apprehensions about this being an exacerbation. An exacerbation is pretty much defined simply as a temporary worsening of the lung function due to an infection or inflammation. Although no formal definition exists, an exacerbation is generally characterized by the following symptoms: 1. Shortness of breath 2. Fatigue 3. Increased cough 4. More productive cough 5. Drop in FEV1 or other markers of the pulmonary function tests*

*The protocol at many hospitals is to admit and treat with IV antibiotics if a patient is exhibiting these symptoms, as the assumption is made that they are suffering a pulmonary exacerbation. I did not believe [Son] was having an exacerbation. While I haven't been collection data for long enough to have caught an exacerbation, I recall him being symptomatic 24/7, and [Son]'s serious symptoms seemed somewhat isolated to sleeping or laying down. I had notes in my tracking that called out the times that these things were happening, and at 3pm he was running around the hospital room and playing without a care in the world - asymptomatic. I thought that [Son] had caught the cold that his siblings had and while he was sleeping/laying, the post nasal drip was throwing him into these coughing fits. His symptoms just didn't fit the bill of his "typical" exacerbation.*

*Based on the data I shared from his EHR, the attending doctor came in to discuss the IV antibiotics that they wanted to put him on. If I was going to agree to IV's, they chose the two that I would have also picked based on the positive response he had to them in the past, but I wasn't ready for IVs. Once you start two weeks of IV antibiotics, it isn't common to stop them until they're complete. I once again pulled out my data and had a discussion with the doctor about waiting to see if the steroid and/or the oral antibiotic that he is on would kick in. Since we were already in the hospital, he agreed to monitor [Son] overnight before starting IV's on two conditions: I would agree to start the IV's if his oxygen dropped any lower (he was at 92) or if he had any more coughing spells overnight. He had started the steroid the night before so had only had 1 dose of both that and the oral antibiotic, usually not quite enough time for [Son] to respond. He was also on inhaled Colistin as we work to eradicate his *Achromobacter*. The benefit of Colistin is that it covers *pseudomonas* and the*

*Cipro he was on covers a number of other bacteria that he has grown in the past. It just seemed to me unlikely that he was having a true exacerbation.*

*So we hooked him up to a Pulse Oximeter and put him to bed. Moments later, a doctor stopped in to tell me that his culture came back positive for Rhinovirus, otherwise known as the common cold. It is known to cause a lot of inflammation and it's not unheard of for kids to have lots of coughing with this virus. Kids who have tracheomalacia, [Son], often do end up in the hospital because their airways are under attack more than they can handle on their own, and many times need a steroid to help reduce the inflammation and help them to get over the hump. For the first night in 3 nights, he slept all night long without coughing. His oxygen levels gradually started to go up on their own overnight. He woke up in the morning refreshed and looking/acting well again. I needed to take him to [Son] that morning because I was starting to reach the point of uncomfortable managing it from home in the case that things would get worse before they got better. But had I given it one more night, one more chance for the steroid to kick in and do it's job, we probably could have avoided the hospital altogether.*

*We were discharged at 10:30am the next morning, just under 24hrs after we had checked into the ER. Because I had been tracking his health so closely, because they were able to access his previous test results, because I felt empowered enough to speak up and express my perspective and desired course of action, and because I had given them the evidence to trust me, we found a mutually agreed upon solution that saved us two weeks in the hospital. It saved [Son] the stress of being in the hospital for 2 weeks. It saved me the trouble of trying to arrange my life for a 2 week hospital stay. It saved the doctors and the hospital time and money. My insurance company wasn't being billed for unnecessary tests and an extended hospital stay.*

*I have data on [Son]. I know the signs and symptoms of a pulmonary exacerbation. I know the difference in his cough frequency or appetite that is associated with the onset of symptoms. I know what has worked for him and what hasn't. Having that data and sharing it with anyone and everyone who might be able to better help him or benefit from it personally is why I track it. I lived my "what if" and I couldn't be more pleased with the way it turned out.*

# **Appendix B**

## **Personal Experiments and MyIBD Catalogs**

This appendix documents the measures, treatments, and experiments produced by the author and other users on the two prototype sites described in Chapter 8.

### **B.1 User Study Experiments**

This section contains summary tables for the 10 treatments and experiments used in the user study described in section 9.2.3. Treatment descriptions are contained inline. In the control group, only the treatment descriptions were shared with the participants.

<b>Title</b>	<b>Improve Sleep with Moderate Exercise</b>
<b>Treatment</b>	Increase in Exercise
<b>Hypothesis</b>	Research shows that exercise, particularly high-intensity exercise, improve the quality of your sleep for many people. We should see a noticeable effect from moderate increase, or very short daily periods of high intensity (stair climbing, etc).
<b>Outcome</b>	Total Sleep
<b>Covariates</b>	Time to sleep, Steps,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	2 / 3
<b>Treatment Description</b>	Double your weekly minutes of exercise, or at least 20 minutes of regular daily fast walking or other cardio exercise. Shorter periods of higher intensity are better, but not required.
<b>Side Effects</b>	If you have a heart condition, or are out of shape this can cause breathlessness or exhaustion. Not recommended for people who are severely out of shape without medical supervision.
<b>Protocol</b>	During the treatment period, make sure you follow the treatment every day in order to see the benefit. During off days, try to maintain a normal activity level for you. We'll also track steps and total sleep in case you are too far off. Total steps measured by the a Jawbone or other activity tracker is a good measure of whether your activity really was greater during the treatment period.



<b>Title</b>	<b>Improve productivity with a buddy.</b>
<b>Treatment</b>	Buddy Up
<b>Hypothesis</b>	Working with peers can be more productive because it helps create a shared context, keeps us on task, etc.
<b>Outcome</b>	Computer Productivity
<b>Covariates</b>	Total time on computers, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Using IM, chat, or Facebook messaging to interact with someone throughout the day who has a shared goal. This should be someone trusted with whom you can share your failures and plot success. Use trackers like Rescuetime (productivity) and Jawbone/Fitbit (running/walking/exercise) and shared your daily data to make it a friendly contest with daily post-mortem.
<b>Side Effects</b>	None. Treatment period should overlap with your buddy if they are also experimenting.
<b>Protocol</b>	<p>Agree with a buddy that you will work together during the study period to keep each other on task. Agree on a way to do it for the duration of the period. For example, you can talk on instant messaging throughout the day, compete in a friendly way through sharing Rescuetime data, etc.</p> <p>Here is an example invitation paragraph you can start with:  ”Dear ;friend;, I’m doing a study on how to improve working habits by connecting with another person who is also working to improve theirs. This would involve you and I talking a few times a day (over IM, e-mail, etc) over one work week so we can help each other stay on task and reflect on our habits. If you are interested, please let me know and we can discuss how we would best work together on this. Thank you, ;me;”</p> <p>Please share your your experiences and hints in the comments tab.</p>

<b>Title</b>	<b>Improve Productivity through "Self Control" and "Get Focused"</b>
<b>Treatment</b>	Self-Control Application
<b>Hypothesis</b>	Using the self control application for 90 minutes or more a day will reduce the number of hours spent looking at e-mail and unproductive websites leaving more time for productive work.
<b>Outcome</b>	Computer Productivity
<b>Covariates</b>	Total time on computers, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Install a "self-control" application that turns off access to email and/or the web for defined periods of time. This includes Rescuetime's Get Focused feature (although it doesn't turn off e-mail). If you really need to see e-mail, you can get it on your phone, but this makes it harder if you constantly check email/facebook during the workday. Works best if you mostly work on the computer.
<b>Side Effects</b>	None known.
<b>Protocol</b>	During the treatment period, use the Self-Control Application for at least 90 minutes at a time and no less than 4.5 hours in a day. If you really need to keep in touch with e-mail, use a mobile device but try not to cheat too much. You can journal if you find that you are trading time on your main machine for time on your mobile.

<b>Title</b>	<b>Improve your productivity with the Pomodoro technique</b>
<b>Treatment</b>	Pomodoro Technique
<b>Hypothesis</b>	The Pomodoro technique is a system for helping maintain focus and build an awareness of how long tasks take by breaking them down into short, well-defined chunks.
<b>Outcome</b>	Computer Productivity
<b>Covariates</b>	Total time on computers,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Use the formal or a variation of the Pomodoro Technique to break your time down into smaller chunks. There are many tools that support this technique.
<b>Side Effects</b>	None known.
<b>Protocol</b>	During the baseline period, read up on the Pomodoro technique. Purchase (if you choose) a program like Vitamin-R. Try out the tool and technique during the treatment period and stop afterwards.

<b>Title</b>	<b>Use f.lux to improve sleep regulation</b>
<b>Treatment</b>	f.lux Application
<b>Hypothesis</b>	F.lux removes blue light from your computer at sundown, improving the natural melatonin cycle disturbed by blue light from computer screens.
<b>Outcome</b>	Total Sleep
<b>Covariates</b>	Time to sleep, Sleep Quality,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	4 / 0
<b>Treatment Description</b>	The f.lux application removes the blue from your computer display after the sun goes down. In theory, this helps restore normal melatonin production and prepare you for sleeping. Watch TV on an f.lux computer if you must watch TV at night. This will happen automatically when you install the app, so you can disable reminders if experimenting with this.
<b>Side Effects</b>	None known.
<b>Protocol</b>	During the treatment period, enable f.lux. During the baseline periods, disable it.

<b>Title</b>	<b>Improve sleep regulation with melatonin</b>
<b>Treatment</b>	Melatonin
<b>Hypothesis</b>	Melatonin is our sleep hormone; small oral doses have been shown to help some people fall asleep who have insomnia induced by a disturbed melatonin cycle.
<b>Outcome</b>	Sleep Quality
<b>Covariates</b>	Time to sleep, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	2 / 2
<b>Treatment Description</b>	An aid for restoring the quality of your sleep. Melatonin is often called the "Sleep Hormone". Supplementation improves the natural increase of blood concentrations of melatonin at night, accelerating recovery from jet lag or reducing the time to fall asleep. Melatonin is only intended for short term use (up to 2 months).
<b>Side Effects</b>	May cause a groggy feeling the next morning. You can reduce your dose to as little as 0.25mg by breaking up 1mg pills to see if you can reduce grogginess. Less common side effects may include abdominal discomfort, mild anxiety, irritability, confusion and short-lasting feelings of depression. Melatonin can also interact with medications. Do not use if taking anti-coagulant, immunosuppressant drugs, diabetes medication or birth control pills. Check with your doctor!
<b>Protocol</b>	Take Melatonin between 2 hours and 30 minutes of bedtime. Be consistent throughout the treatment period. Dose should be 0.5-1.0 mg (split a 1mg tablet if you need to).

<b>Title</b>	<b>Improve bedtime with amber glasses</b>
<b>Treatment</b>	Amber Glasses
<b>Hypothesis</b>	Amber glasses decrease the amount of blue light hitting our retina. This should help regulate the melatonin cycle and make us naturally more sleep, especially if we use the computer or watch TV at night.
<b>Outcome</b>	Total Sleep
<b>Covariates</b>	Time to sleep, Sleep Quality,
<b>Design</b>	null
<b>Onset/washout</b>	4 / 4
<b>Treatment Description</b>	Purchase a pair of amber glasses. Wear them from sundown until bedtime. Should help to improve the melatonin cycle. You might look weird, but you can watch TV with these vs. using things like f.lux and lowering house lights.
<b>Side Effects</b>	None known.
<b>Protocol</b>	During the treatment period, wear amber glasses starting 2-3 hours before bed, or at sundown if you go to sleep late.

<b>Title</b>	<b>Improve productivity with multiple behaviors</b>
<b>Treatment</b>	Productivity Collection
<b>Hypothesis</b>	If one approach can help, why not all?
<b>Outcome</b>	Computer Productivity
<b>Covariates</b>	Total time on computers, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Sleep 8 hours a night, install a self-control application, use the Pomodoro technique, and share your productivity data with a buddy.
<b>Side Effects</b>	None.
<b>Protocol</b>	During the treatment period: - Work with a buddy to stay on track - Use "Self Control" and "Stay focused" to avoid email and web distractions - Use the Pomodoro technique to break up your work - Ensure you get near to 8 hours of sleep a night!

<b>Title</b>	<b>White noise decreases time to fall asleep</b>
<b>Treatment</b>	White Noise in Bedroom
<b>Hypothesis</b>	If you are easily disturbed by sleep noises, white noise can raise our perceptual "noise floor", increasing the level of noise needed to disturb our sleep, making it more likely we'll fall and stay asleep.
<b>Outcome</b>	Time to Fall Asleep
<b>Covariates</b>	Sleep Quality, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	You can use a physical fan, a variety of white noise or other ambient sound generator (there is an app for that), or any other source of low-grade background noise. This reduces sleep disturbance due to intermittent outside sounds by raising the threshold.
<b>Side Effects</b>	Ear problems if you make it too loud!
<b>Protocol</b>	Turn on a source of white noise before going to bed each night during the treatment period. Make it as loud as you can comfortably ignore. You can sign up for treatment reminders and set them for your desired bedtime.

<b>Title</b>	<b>Boost time-on-task productivity through sleep</b>
<b>Treatment</b>	Sleep 8 hours a night
<b>Hypothesis</b>	Most people can improve their 'time on task' productivity by increasing their nightly sleep which improves memory, attention, etc.
<b>Outcome</b>	Computer Productivity
<b>Covariates</b>	Total time on computers, Total Sleep,
<b>Design</b>	Period-length: 7, Randomized: no
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Force yourself to sleep 8 hours a night for at least 4 days. Great intervention to see what the impact of sleep can be on mood, productivity, time to sleep, etc.
<b>Side Effects</b>	Feeling awesome!
<b>Protocol</b>	During the treatment period, work hard to get 8 hours of sleep a night. The period is kept short to make this easy!

## B.2 Personal Experiments Catalog

### B.2.1 Experiments

<b>Title</b>	<b>Can oral Rifaximin reduce abdominal pain?</b>
<b>Treatment</b>	Rifaximin
<b>Hypothesis</b>	Rifaximin will reduce the amount of bad bacteria and decrease gas production and bloating.
<b>Outcome</b>	Pain
<b>Covariates</b>	PROMIS Pain, PROMIS Fatigue,
<b>Design</b>	Period-length: 14, Randomized: no
<b>Onset/washout</b>	7 / 3
<b>Treatment Description</b>	An antibiotic that is not absorbed by the body and only treats the GI tract.
<b>Side Effects</b>	Most people experience only limited side effects, but some unpleasant GI symptoms and other side effects are experienced by some. See the [Drugs.com Overview]( <a href="http://www.drugs.com/sfx/rifaximin-side-effects.html">http://www.drugs.com/sfx/rifaximin-side-effects.html</a> )
<b>Protocol</b>	Ask your doctor to prescribe a course of rifaximin for at least 3 weeks, the system will let you know when it is time to start taking it.

<b>Title</b>	<b>Increase efficiency through self control</b>
<b>Treatment</b>	Self-Control Application
<b>Hypothesis</b>	Self control/awareness will improve computer work efficiency.
<b>Outcome</b>	Computer Work Efficiency
<b>Covariates</b>	Steps, Mood Score, Total Sleep,
<b>Design</b>	null
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	Install a "self-control" application that turns off access to email and/or the web for defined periods of time. This includes Rescuetime's Get Focused feature (although it doesn't turn off e-mail). If you really need to see e-mail, you can get it on your phone, but this makes it harder if you constantly check email/facebook during the workday. Works best if you mostly work on the computer.
<b>Side Effects</b>	None known.
<b>Protocol</b>	Three times throughout the day switch to self control for 45 minutes at a time. These reminders will be sent through email and SMS at 930AM, 130PM, and 430PM. At this time turn on self control (on mac) or close email/social media/news sites (PC or Mac) or use a pre-created blacklist filter (various programs/apps). Mood Score will be tracked through SMS.

<b>Title</b>	<b>Daily exercise to improve quality of sleep</b>
<b>Treatment</b>	Increase in Exercise
<b>Hypothesis</b>	I think that daily exercise will decrease stress. So my quality of sleep will improve
<b>Outcome</b>	Sleep Quality
<b>Covariates</b>	Time to sleep,
<b>Design</b>	null
<b>Onset/washout</b>	2 / 3
<b>Treatment Description</b>	Double your weekly minutes of exercise, or at least 20 minutes of regular daily fast walking or other cardio exercise. Shorter periods of higher intensity are better, but not required.
<b>Side Effects</b>	If you have a heart condition, or are out of shape this can cause breathlessness or exhaustion. Not recommended for people who are severely out of shape without medical supervision.
<b>Protocol</b>	I will track my exercises with Suunto Ambit and maybe I need some tracker to measure my sleep.(Fitbit)

<b>Title</b>	<b>Does daily exercise improve my energy level?</b>
<b>Treatment</b>	Increase in Exercise
<b>Hypothesis</b>	Daily exercise will help to improve the quality of my sleep, thus leaving me better rested and more energetic
<b>Outcome</b>	Energy Level
<b>Covariates</b>	Total Sleep,
<b>Design</b>	null
<b>Onset/washout</b>	2 / 3
<b>Treatment Description</b>	Double your weekly minutes of exercise, or at least 20 minutes of regular daily fast walking or other cardio exercise. Shorter periods of higher intensity are better, but not required.
<b>Side Effects</b>	If you have a heart condition, or are out of shape this can cause breathlessness or exhaustion. Not recommended for people who are severely out of shape without medical supervision.
<b>Protocol</b>	I will walk 30 minutes at least 5 days a week. In the evening, I will complete a mood score to determine if exercise has an impact on my mood and energy level.

<b>Title</b>	<b>Assessing the effect of 500mg L-Tryptophan before bed time on hours of deep sleep.</b>
<b>Treatment</b>	L-Tryptopan
<b>Hypothesis</b>	It's a precursor for serotonin which is converted into melatonin. There may be other mechanism. I'll look into this more.
<b>Outcome</b>	Deep Sleep
<b>Covariates</b>	
<b>Design</b>	null
<b>Onset/washout</b>	1 / 89
<b>Treatment Description</b>	Take 500mg L-Tryptophan 15 minutes before bedtime. The goal for bed time is 12.15am. The L-Tryptophan should be taken at 12am. NOTE: onset and washout are just guesses. I need to research that.
<b>Side Effects</b>	None known.
<b>Protocol</b>	I'll start taking L-Tryptophan for six nights (Tonight and next week; excluding Friday and Saturday). After that I'll quit taking L-Tryptophan for five nights. At that point I'll look compare the percentage and total amount of deep sleep during the treatment and control days. I will also compare it to the large sleep database I already have to see if L-Tryptophan has an effect.



<b>Title</b>	<b>Glycerin and Witch Hazel for Psoriasis</b>
<b>Treatment</b>	Glycerin and Witch Hazel
<b>Hypothesis</b>	Topical or oral Glycerin and Witch Hazel help re-establish normal barriers between layers of developing skin cells that are disrupted by the inflammatory processes in psoriasis patients.
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Itchiness, Psoriatic Redness,
<b>Design</b>	Period-length: 3, Randomized: no
<b>Onset/washout</b>	21 / 21
<b>Treatment Description</b>	<p>A topical treatment some patients find help with symptoms of psoriasis. Mix glycerin 50/50 with alcohol-free and scent-free witch hazel. Apply as skin feels dry or itchy, but at least every morning and evening. A spray bottle can ease application. After applying, rub in until solution disappears and skin remains moist. If it is greasy, then too much has been applied.</p> <p>Impact on scaling in psoriasis can be seen in responding patients within two weeks, often sooner for other symptoms such as redness and scaling.</p>
<b>Side Effects</b>	
<b>Protocol</b>	<p>Choose one area of your body to treat for this study, treating the rest of your body as you normally do until the end of the trial. This will help you compare and contrast your normal habits with the treatment.</p> <p>Start by measuring normal psoriasis activity for a week, apply the treatment for 3 weeks, then remove the treatment again to see if symptoms return without treatment for another two weeks.</p> <p><b>WARNING:</b> For some people symptoms do not return, in which case the trial may report a negative outcome in error when in fact it was an unqualified success. The trial may be updated in the future to account for this.</p>

<b>Title</b>	<b>Use white noise at night to reduce awakenings</b>
<b>Treatment</b>	White Noise in Bedroom
<b>Hypothesis</b>	White noise raises the auditory noise floor, making us less susceptible to sleep disturbance caused by random outside or inside noises.
<b>Outcome</b>	Nightly Awakenings
<b>Covariates</b>	Total Sleep,
<b>Design</b>	null
<b>Onset/washout</b>	0 / 0
<b>Treatment Description</b>	You can use a physical fan, a variety of white noise or other ambient sound generator (there is an app for that), or any other source of low-grade background noise. This reduces sleep disturbance due to intermittent outside sounds by raising the threshold.
<b>Side Effects</b>	Ear problems if you make it too loud!
<b>Protocol</b>	During the treatment period, add a fan or other source of white noise to your room as loud as you can tolerate and still fall asleep. You can reduce the noise level in the future and tack your awakenings.

<b>Title</b>	<b>Mood + Exercise</b>
<b>Treatment</b>	Increase in Exercise
<b>Hypothesis</b>	The increase in exercise should improve mood, and as well, an increase in mood should improve ability/desire to exercise.
<b>Outcome</b>	Mood
<b>Covariates</b>	
<b>Design</b>	null
<b>Onset/washout</b>	2 / 3
<b>Treatment Description</b>	Double your weekly minutes of exercise, or at least 20 minutes of regular daily fast walking or other cardio exercise. Shorter periods of higher intensity are better, but not required.
<b>Side Effects</b>	If you have a heart condition, or are out of shape this can cause breathlessness or exhaustion. Not recommended for people who are severely out of shape without medical supervision.
<b>Protocol</b>	Week One: Track Mood in several instances each day Week Two: Track Exercise throughout the week Week Three: Track both Mood and Exercise daily/throughout the week.

<b>Title</b>	<b>Medication adherence to reduce pain/discomfort</b>
<b>Treatment</b>	Adherence to Medications
<b>Hypothesis</b>	Medication that is intended to reduce pain (e.g. GI pain) only has an effect if taken regularly; use the treatment reminders to compare a reminder-based adherence strategy to a non-reminder based one.
<b>Outcome</b>	Pain
<b>Covariates</b>	mood score,
<b>Design</b>	Period-length: 10, Randomized: no
<b>Onset/washout</b>	3 / 3
<b>Treatment Description</b>	See if I feel better overall if I adhere better, use the "how I feel" measurement to double as a reminder at the various times of day I need to be prompted.
<b>Side Effects</b>	None known
<b>Protocol</b>	When you get a treatment reminder, lookup what medications you should take and take them. Also setup once or twice daily assessments of mood and pain scores to see if there is an improvement.

<b>Title</b>	<b>Curcumin for Psoriasis</b>
<b>Treatment</b>	Curcumin (Turmeric)
<b>Hypothesis</b>	Curcumin, the active ingredient in turmeric is effective in slowing down the energy supply to the rapidly dividing cells which result in red, scaly, skin symptoms. Taken as a dietary supplement, Turmeric or Curcumin can help to improve the appearance of red, scaly, inflamed skin
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Itchiness, Psoriatic Redness,
<b>Design</b>	null
<b>Onset/washout</b>	21 / 10
<b>Treatment Description</b>	<p>Turmeric (<i>Curcuma longa</i>), the major ingredient of curry powder and prepared mustard, has a long history in both Chinese and Ayurvedic (Indian) medicine as an anti-inflammatory agent.</p> <p>The volatile oil fraction of turmeric has demonstrated potent anti-inflammatory activity in a variety of experimental animal models, while curcumin, the yellow pigment of turmeric is even more potent in acute inflammation</p> <ol style="list-style-type: none"> <li>1. When used orally, curcumin inhibits leukotriene formation, inhibits platelet aggregation and stabilizes neutrophilic lysosomal membranes, thus inhibiting inflammation at the cellular level</li> <li>2. Curcumin is reported to possess greater anti-inflammatory activity than ibuprofen</li> <li>3. At low levels, curcumin is a prostaglandin inhibitor, while at higher levels it stimulates the adrenal glands to secrete cortisone</li> <li>4. Formulation difficulties due to the yellow color of curcumin has made topical use slow in coming. However, recent developments in technology may change that. The standard oral dose of curcumin is 250-400 mg, three times a day.</li> </ol>
<b>Side Effects</b>	<p>- Turmeric has been used for centuries as a medicine and in food by Asians. They also use turmeric to treat minor cuts, wounds and burns. This has found to be extremely effective. However there is not enough research to establish the risks of very high intake of turmeric and hence it is best to stick to the recommended dose. - Turmeric is known to act as a blood thinner or anti coagulant. Hence this should not be taken with other blood thinning medicines. Some examples of blood thinning medicine are aspirin, clopodogrel, enoxaparin, heparin etc. - Turmeric should also be avoided by those who have congestive heart disease, gallstones, liver disease, jaundice or acute bilious colic. - In rare cases turmeric could cause diarrhea, sweating and nausea. - Some patients report an initial flare of symptoms followed by improvement</p>
<b>Protocol</b>	The easiest option to get the required dosage of turmeric is through dietary supplements. The suggested dosage of turmeric for psoriasis is 500mg of curcumin three times a day.

<b>Title</b>	<b>Elimination Diet for Psoriasis</b>
<b>Treatment</b>	Minimal Diet for Psoriasis
<b>Hypothesis</b>	Auto-immune diseases often respond to changes in diet, particularly those that are anti-inflammatory. A more aggressive elimination diet is often a good way to test whether dietary changes will work for you at all as a first step into investigating lifestyle therapies. You should see noticeable improvement during the treatment phase of the trial, enough to motivate a longer elimination or trying out other diets.
<b>Outcome</b>	Psoriatic Scaling
<b>Covariates</b>	Psoriatic Itchiness, Psoriatic Redness,
<b>Design</b>	null
<b>Onset/washout</b>	18 / 4
<b>Treatment Description</b>	<p>This diet minimizes sugar and other potentially triggers of immune response such as gluten, dairy, and nightshades. It has been used successfully by some patients to reduce the severity of psoriasis.</p> <p>The elimination diet guidelines include:</p> <ul style="list-style-type: none"> <li>- Gluten free (no wheat-based products, focus on alternative grains) -</li> <li>Low-sugar (no processed sugar, severely limit milk, limit fruits to 1-2 servings a day) -</li> <li>Remove common trigger foods: chocolate, strawberries, tomatoes, and other nightshades. -</li> <li>Limit consumption of red and processed meats (1-2 servings / week) -</li> <li>Supplementation with Omega-3 oil and/or increased consumption of fatty fish (salmon, mackerel, sardines)</li> </ul> <p>Examples of food that can be eaten:</p> <ul style="list-style-type: none"> <li>- Poultry, fish, and eggs -</li> <li>Berries in moderate quantities are acceptable -</li> <li>Any vegetables not on the trigger list -</li> <li>Pastas and cereals made from potato or rice flours -</li> <li>Use unsweetened Almond Milk instead of Milk for cereal, shakes -</li> <li>Low-lactose milk-based products such as hard cheeses are fine -</li> <li>Healthy fats such as olive oil or flax seed oils for flavor -</li> <li>All nuts -</li> <li>Large salads with beans, chopped veggies, and oil-based dressings are great; can add canned salmon or chicken for protein</li> </ul>
<b>Side Effects</b>	null
<b>Protocol</b>	All things in moderation. Minor violations of the diet will not invalidate the results, but they will take longer to establish whether or not you will respond.

<b>Title</b>	<b>Tea Tree Oil for Tinnea Versicolor</b>
<b>Treatment</b>	Tea Tree Oil
<b>Hypothesis</b>	Tea tree oil as anti-bacterial properties. Some people report that it has helped to decrease the symptoms of Tinnea Versicolor.
<b>Outcome</b>	Itching
<b>Covariates</b>	
<b>Design</b>	null
<b>Onset/washout</b>	2 / 5
<b>Treatment Description</b>	<p>Tea tree oil, or melaleuca oil, is a pale yellow color to nearly colorless and clear essential oil with a fresh camphoraceous odor. It is taken from the leaves of the <i>Melaleuca alternifolia</i>, which is native to Southeast Queensland and the Northeast coast of New South Wales, Australia. Tea tree oil should not be confused with tea oil, the sweet seasoning and cooking oil from pressed seeds of the tea plant <i>Camellia sinensis</i> (beverage tea), or the tea oil plant <i>Camellia oleifera</i>.</p> <p>Tea tree oil has been scientifically investigated only recently. Some sources suggest beneficial medical properties when applied topically, including antiviral,[7] antibacterial, antifungal, and antiseptic qualities. (From Wikipedia)</p>
<b>Side Effects</b>	<p>According to the American Cancer Society: "Tea tree oil is toxic when swallowed. It has been reported to cause drowsiness, confusion, hallucinations, coma, unsteadiness, weakness, vomiting, diarrhea, stomach upset, blood cell abnormalities, and severe rashes. It should be kept away from pets and children."</p> <p>Some people can experience allergic contact dermatitis as a reaction to dermal contact with tea tree oil. Allergic reactions may be due to the various oxidation products that are formed by exposure of the oil to light and/or air.</p> <p>If used in concentrations below 4% or particularly below 0.25%, tea tree oil may fail to kill bacteria and create selection pressure, which may result in them becoming less sensitive to tea tree oil and even some antibiotics in vitro.</p>
<b>Protocol</b>	Swab a dilute concentration of tea tree oil (ideally, 4% but more is OK if you can tolerate) on the effected areas once daily during the treatment period.

<b>Title</b>	<b>Test - Sleep Experiment</b>
<b>Treatment</b>	Sleep 8 hours a night
<b>Hypothesis</b>	going to bed early will improve mood
<b>Outcome</b>	Total Sleep
<b>Covariates</b>	Mood,
<b>Design</b>	null
<b>Onset/washout</b>	4 / 2
<b>Treatment Description</b>	Force yourself to sleep 8 hours a night for at least 4 days. Great intervention to see what the impact of sleep can be on mood, productivity, time to sleep, etc.
<b>Side Effects</b>	Feeling awesome!
<b>Protocol</b>	kggkkgkgk

<b>Title</b>	<b>How does daily exercise affect my energy level the following day</b>
<b>Treatment</b>	Increase in Exercise
<b>Hypothesis</b>	Daily exercise will improve the quality of my sleep and thus result in increased energy level the following day.
<b>Outcome</b>	Energy Level
<b>Covariates</b>	Sleep Quality,
<b>Design</b>	null
<b>Onset/washout</b>	2 / 3
<b>Treatment Description</b>	Double your weekly minutes of exercise, or at least 20 minutes of regular daily fast walking or other cardio exercise. Shorter periods of higher intensity are better, but not required.
<b>Side Effects</b>	If you have a heart condition, or are out of shape this can cause breathlessness or exhaustion. Not recommended for people who are severely out of shape without medical supervision.
<b>Protocol</b>	

<b>Title</b>	<b>Protein/Fat Diet</b>
<b>Treatment</b>	Protein/Fat Meals
<b>Hypothesis</b>	Cut Carbs, Cut Sugar, Cut Calorie intake.
<b>Outcome</b>	Weight
<b>Covariates</b>	
<b>Design</b>	null
<b>Onset/washout</b>	3 / 1
<b>Treatment Description</b>	50g Protein Powder 60g Cashews Every 3 Hours Does this make you lose weight?
<b>Side Effects</b>	Lack of vitamins/minerals Lack of variety
<b>Protocol</b>	Eat Protein Powder (50g) and Cashews (60g) every 3 hours as long as you are awake. Measure weight once a day.

<b>Title</b>	<b>Tea Tree Oil for Itching</b>
<b>Treatment</b>	Tea Tree Oil
<b>Hypothesis</b>	Tea tree oil is a natural antifungal and tinnea versicolor is a fungus
<b>Outcome</b>	Itching
<b>Covariates</b>	Deep Sleep, Mood,
<b>Design</b>	null
<b>Onset/washout</b>	2 / 5
<b>Treatment Description</b>	<p>Tea tree oil, or melaleuca oil, is a pale yellow color to nearly colorless and clear essential oil with a fresh camphoraceous odor. It is taken from the leaves of the <i>Melaleuca alternifolia</i>, which is native to Southeast Queensland and the Northeast coast of New South Wales, Australia. Tea tree oil should not be confused with tea oil, the sweet seasoning and cooking oil from pressed seeds of the tea plant <i>Camellia sinensis</i> (beverage tea), or the tea oil plant <i>Camellia oleifera</i>.</p> <p>Tea tree oil has been scientifically investigated only recently. Some sources suggest beneficial medical properties when applied topically, including antiviral,[7] antibacterial, antifungal, and antiseptic qualities. (From Wikipedia)</p>
<b>Side Effects</b>	<p>According to the American Cancer Society: "Tea tree oil is toxic when swallowed. It has been reported to cause drowsiness, confusion, hallucinations, coma, unsteadiness, weakness, vomiting, diarrhea, stomach upset, blood cell abnormalities, and severe rashes. It should be kept away from pets and children."</p> <p>Some people can experience allergic contact dermatitis as a reaction to dermal contact with tea tree oil. Allergic reactions may be due to the various oxidation products that are formed by exposure of the oil to light and/or air.</p> <p>If used in concentrations below 4% or particularly below 0.25%, tea tree oil may fail to kill bacteria and create selection pressure, which may result in them becoming less sensitive to tea tree oil and even some antibiotics in vitro.</p>
<b>Protocol</b>	<p>Establish itching baseline. Apply Tea Tree Oil treatment as described. Should see effect in about 3 days. Goal is zero itching.</p> <p>Side effects sometimes include skin drying. Apply moisturizer as needed.</p>



<b>Title</b>	<b>Improved sleep with Melatonin</b>
<b>Treatment</b>	Melatonin
<b>Hypothesis</b>	Melatonin will help improve quality of sleep and help me stay asleep longer.
<b>Outcome</b>	Quality of Sleep
<b>Covariates</b>	
<b>Design</b>	null
<b>Onset/washout</b>	2 / 1
<b>Treatment Description</b>	An aid for restoring the quality of your sleep. Melatonin is often called the "Sleep Hormone". Supplementation improves the natural increase of blood concentrations of melatonin at night, accelerating recovery from jet lag or reducing the time to fall asleep. Melatonin is only intended for short term use (up to 2 months).
<b>Side Effects</b>	May cause a groggy feeling the next morning. You can reduce your dose to as little as 0.25mg by breaking up 1mg pills to see if you can reduce grogginess. Less common side effects may include abdominal discomfort, mild anxiety, irritability, confusion and short-lasting feelings of depression. Melatonin can also interact with medications. Do not use if taking anti-coagulant, immunosuppressant drugs, diabetes medication or birth control pills. Check with your doctor!
<b>Protocol</b>	Take 1 mg of Melatonin within one hour of bedtime; use Jawbone Up to track progress

<b>Title</b>	<b>Test Experiment</b>
<b>Treatment</b>	Sleep 8 hours a night
<b>Hypothesis</b>	Increased Sleep time will lead to better focus during the day
<b>Outcome</b>	Fitbit Activity Score
<b>Covariates</b>	Mood Score,
<b>Design</b>	null
<b>Onset/washout</b>	4 / 2
<b>Treatment Description</b>	Force yourself to sleep 8 hours a night for at least 4 days. Great intervention to see what the impact of sleep can be on mood, productivity, time to sleep, etc.
<b>Side Effects</b>	Feeling awesome!
<b>Protocol</b>	Go to bed early

## B.2.2 Instruments

<b>Variable</b>	<b>Deep Sleep</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Deep sleep measured by Jawbone UP. This is measured based on activity level, but does not indicate true restorative deep sleep like Zeo or an EEG-based sleep tracker would
<b>Origin</b>	System

<b>Variable</b>	<b>Itching</b>
<b>Service</b>	SMS
<b>Description</b>	Rate your itching 0-3 (0=none, 3=bad)
<b>Origin</b>	User

<b>Variable</b>	<b>Psoriatic Scaling</b>
<b>Service</b>	SMS
<b>Description</b>	A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. The most important thing is to try to be consistent in what the psoriasis looks like from measure to measure.
<b>Origin</b>	System

<b>Variable</b>	<b>Time Awake</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Time between button press and getting up measured by Jawbone UP where user was not sleeping
<b>Origin</b>	System

<b>Variable</b>	<b>Sleep Efficiency</b>
<b>Service</b>	FitBit
<b>Description</b>	The % of time you were asleep while in bed
<b>Origin</b>	System

<b>Variable</b>	<b>Time to sleep</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Time in minutes between when you pressed the Jawbone button and when it recorded that you fell asleep. Press the button at a consistent time in your bedtime routine, such as when you get into bed, when you close the book/computer, turn out the light, stop talking to your SO, etc.
<b>Origin</b>	System

<b>Variable</b>	<b>Distance</b>
<b>Service</b>	FitBit
<b>Description</b>	Total daily distance traveled according to fitbit
<b>Origin</b>	System

<b>Variable</b>	<b>Total Bed Time</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Total time in bed, both awake and asleep. Measured from button press to button press or moment of increased activity
<b>Origin</b>	System

<b>Variable</b>	<b>Fitbit Activity Score</b>
<b>Service</b>	FitBit
<b>Description</b>	Fitbit's activity level rollup score
<b>Origin</b>	System

<b>Variable</b>	<b>Total Sleep</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Total hours of sleep according to the Jawbone UP
<b>Origin</b>	System

<b>Variable</b>	<b>Total Calories</b>
<b>Service</b>	FitBit
<b>Description</b>	Total daily calories expended, activities + estimated
<b>Origin</b>	System

<b>Variable</b>	<b>Computer Entertainment</b>
<b>Service</b>	RescueTime
<b>Description</b>	Your total entertainment hours online as measured by Rescuetime
<b>Origin</b>	System

<b>Variable</b>	<b>Light Sleep</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Light sleep measured by Jawbone UP. This is not light sleep like you would get from an EEG like zeo, but basically periods of increased physical motion
<b>Origin</b>	System

<b>Variable</b>	<b>Well Being</b>
<b>Service</b>	SMS
<b>Description</b>	A simple well-being score that can be used to summarize the last day or week ranging from 0 = poor to 5 = fair to 10 = well/normal
<b>Origin</b>	System

<b>Variable</b>	<b>Total Sleep</b>
<b>Service</b>	FitBit
<b>Description</b>	Total hours asleep, not including waking time
<b>Origin</b>	System

<b>Variable</b>	<b>Oxygen Saturation</b>
<b>Service</b>	SMS
<b>Description</b>	Room air vs O2
<b>Origin</b>	User

<b>Variable</b>	<b>Energy Level</b>
<b>Service</b>	SMS
<b>Description</b>	On a scale of 1-5 with 1 being totally exhausted and 5 being totally energized, how do you feel?
<b>Origin</b>	User

<b>Variable</b>	<b>Steps</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Number of steps in a day according to Jawbone UP v2
<b>Origin</b>	System

<b>Variable</b>	<b>Sleep Quality</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Sleep quality score by Jawbone UP
<b>Origin</b>	System

<b>Variable</b>	<b>Nightly Awakenings</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Number of times Jawbone activity level indicated I was awake or in high REM
<b>Origin</b>	System

<b>Variable</b>	<b>Total Bed Time</b>
<b>Service</b>	FitBit
<b>Description</b>	The total time you spent in bed before heading to seel and getting up for the day.
<b>Origin</b>	System

<b>Variable</b>	<b>Time to Fall Asleep</b>
<b>Service</b>	SMS
<b>Description</b>	The time in minutes from when you go to bed attempting to go to sleep until you are actually asleep (as measured by a device like Fitbit, Up, or Zeo). This is a manual measurement, but perhaps we can extract this measure from various devices in the future?
<b>Origin</b>	User

<b>Variable</b>	<b>Nightly Awakenings</b>
<b>Service</b>	FitBit
<b>Description</b>	Count of the number of times you work up between going to sleep and getting up
<b>Origin</b>	System

<b>Variable</b>	<b>Time Awake</b>
<b>Service</b>	FitBit
<b>Description</b>	Total hours awake during sleeping hours
<b>Origin</b>	System

<b>Variable</b>	<b>Psoriatic Itchiness</b>
<b>Service</b>	SMS
<b>Description</b>	A simple measure of the degree of active scaling in a region of psoriasis. The region should be distinct and about the size of your hand. Using the right numbers is less important than using the same numbers for the same degree of symptom from day to day.
<b>Origin</b>	System

<b>Variable</b>	<b>Time to sleep</b>
<b>Service</b>	FitBit
<b>Description</b>	Time to fall asleep in minutes
<b>Origin</b>	System

<b>Variable</b>	<b>Psoriatic Redness</b>
<b>Service</b>	SMS
<b>Description</b>	A measure of the amount of redness in a chosen region of psoriasis. The region should be distinct and about the size of your hand. Try to use the same numbers for the same degree of symptom from day to day.
<b>Origin</b>	System

<b>Variable</b>	<b>Steps</b>
<b>Service</b>	FitBit
<b>Description</b>	Number of steps in a day according to fitbit
<b>Origin</b>	System

<b>Variable</b>	<b>Deep Sleep Ratio</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Deep sleep time divided by total sleep
<b>Origin</b>	System

<b>Variable</b>	<b>Rushing to the Bathroom</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	System

<b>Variable</b>	<b>Stools with Blood</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	System

<b>Variable</b>	<b>Mood</b>
<b>Service</b>	SMS
<b>Description</b>	What is your mood today
<b>Origin</b>	User

<b>Variable</b>	<b>Quality of Sleep</b>
<b>Service</b>	SMS
<b>Description</b>	Tracking the quality of sleep. 0 - Less than 2 hours 1 - 2-3 hours 2- 3-5 hours 3 - 5-6 hours or 6 + hours with 2 interruptions 4 - 6+ hours with 1 interruption 5 - 6+ hours with 0 interruptions
<b>Origin</b>	System

<b>Variable</b>	<b>Mood Score</b>
<b>Service</b>	SMS
<b>Description</b>	A simple assessment of your mood. The specific numbers don't matter, just that you use similar numbers for similar moods.
<b>Origin</b>	System

<b>Variable</b>	<b>PROMIS Pain</b>
<b>Service</b>	Survey
<b>Description</b>	PROMIS Pain weekly measurement.
<b>Origin</b>	System

<b>Variable</b>	<b>Heart Rate</b>
<b>Service</b>	Withings
<b>Description</b>	Resting heart rate during a blood pressure reading
<b>Origin</b>	System

<b>Variable</b>	<b>Feeding tube</b>
<b>Service</b>	SMS
<b>Description</b>	Did you use a feeding tube last night?
<b>Origin</b>	User

<b>Variable</b>	<b>Quality of Sleep</b>
<b>Service</b>	SMS
<b>Description</b>	Time Spent in N-of-1 trials
<b>Origin</b>	System

<b>Variable</b>	<b>Diastolic BP</b>
<b>Service</b>	Withings
<b>Description</b>	Diastolic blood pressure from withings cuff
<b>Origin</b>	System

<b>Variable</b>	<b>PedsQL</b>
<b>Service</b>	Survey
<b>Description</b>	PedsQL Monthly Quality of Life assessment
<b>Origin</b>	System

<b>Variable</b>	<b>PROMIS Fatigue</b>
<b>Service</b>	Survey
<b>Description</b>	PROMIS Fatigue, a weekly measure of overall fatigue commonly used in Pediatric research and patient quality of life assessment.
<b>Origin</b>	System



<b>Variable</b>	<b>Weight</b>
<b>Service</b>	SMS
<b>Description</b>	Measuring body mass.
<b>Origin</b>	User

<b>Variable</b>	<b>Systolic BP</b>
<b>Service</b>	Withings
<b>Description</b>	Systolic blood pressure from withings cuff
<b>Origin</b>	System

<b>Variable</b>	<b>Steps</b>
<b>Service</b>	FitBit
<b>Description</b>	Number of steps in a day according to fitbit
<b>Origin</b>	System

<b>Variable</b>	<b>Sleep Duration</b>
<b>Service</b>	SMS
<b>Description</b>	A manual report of the number of hours of sleep you had last night
<b>Origin</b>	System

<b>Variable</b>	<b>Pain</b>
<b>Service</b>	SMS
<b>Description</b>	A standard metric of pain from 0 to 10 where 0 is no pain and 10 is the worst you've ever felt.
<b>Origin</b>	System

<b>Variable</b>	<b>Daily Stools</b>
<b>Service</b>	SMS
<b>Description</b>	Total stools in the past 24 hours. Best to setup reminders for early in the morning right after you wake up. Keeping a piece of graph paper by the bed at night to make check marks is helpful as well. Write down your daily stools and then provide the total of daily + nightly the next morning.
<b>Origin</b>	System

<b>Variable</b>	<b>Nightly Stools</b>
<b>Service</b>	SMS
<b>Description</b>	Self explanatory. Best to setup reminders for early in the morning right after you wake up. Keeping a piece of graph paper by the bed at night to make check marks is helpful as well.
<b>Origin</b>	System

<b>Variable</b>	<b>Stool consistency</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	System

<b>Variable</b>	<b>Probiotic Use</b>
<b>Service</b>	SMS
<b>Description</b>	Best to be sent just before or just after bed. Remark on intake of probiotic supplements, yogurt, kiefer, probiotic drinks, kombucha, etc.
<b>Origin</b>	System

<b>Variable</b>	<b>Mood</b>
<b>Service</b>	Jawbone UP
<b>Description</b>	Your mood from 0 to 7 according to Jawbone UP
<b>Origin</b>	System

<b>Variable</b>	<b>Calories Consumed</b>
<b>Service</b>	SMS
<b>Description</b>	How many calories did you take in since your last report? Use the note feature to record information by meal, such as '350 breakfast', '400 lunch', '750 dinner'. The note will be visible if you mouse-over a link on the Charts page.
<b>Origin</b>	User

<b>Variable</b>	<b>Fat Mass</b>
<b>Service</b>	Withings
<b>Description</b>	Fat Mass according to Withings Scale
<b>Origin</b>	System

<b>Variable</b>	<b>Airway Clearance</b>
<b>Service</b>	SMS
<b>Description</b>	Tracking the number of Vest treatments Drew does daily.
<b>Origin</b>	User

<b>Variable</b>	<b>Stool Frequency</b>
<b>Service</b>	SMS
<b>Description</b>	Track how many bowel movements have you had in a day as well as their consistency
<b>Origin</b>	User

<b>Variable</b>	<b>LBM</b>
<b>Service</b>	Withings
<b>Description</b>	Lean Body Mass according to Withings
<b>Origin</b>	System

<b>Variable</b>	<b>Sleep duration</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	User

<b>Variable</b>	<b>Fat Ratio</b>
<b>Service</b>	Withings
<b>Description</b>	
<b>Origin</b>	System

<b>Variable</b>	<b>Weight</b>
<b>Service</b>	Withings
<b>Description</b>	Weight as measured by the withings scale
<b>Origin</b>	System

<b>Variable</b>	<b>Height</b>
<b>Service</b>	Withings
<b>Description</b>	
<b>Origin</b>	System

<b>Variable</b>	<b>Computer Work Efficiency</b>
<b>Service</b>	RescueTime
<b>Description</b>	Your computer-based work efficiency as measured by rescuetime
<b>Origin</b>	System

<b>Variable</b>	<b>Exercise</b>
<b>Service</b>	SMS
<b>Description</b>	How many minutes of exercise did you do today?
<b>Origin</b>	User

<b>Variable</b>	<b>Medication Baseline Variation</b>
<b>Service</b>	SMS
<b>Description</b>	Regular medication treatment regimen includes inhaled Albuterol or Atrovent 3x daily, Pulmicort 2x daily, Hypertonic Saline 2x daily, Pulmozyme 1x daily. Oral Prevacid 2x daily, Vitamax 1x daily, Vitamin D 3x per week, Azithromycin 3x per week. Saline nasal rinse and Nasonex spray 2x daily. Miralax 1/2 cap 2x daily. Enter 1 for Normal Routine as listed above Enter 2 if oral antibiotic included Enter 3 if inhaled antibiotic included Enter 4 if IV antibiotic included
<b>Origin</b>	User

<b>Variable</b>	<b>Total time on computers</b>
<b>Service</b>	RescueTime
<b>Description</b>	Total time on a rescuetime enabled computer
<b>Origin</b>	System

<b>Variable</b>	<b>Oxygen Saturation</b>
<b>Service</b>	SMS
<b>Description</b>	Daily oxygen saturation measured with a Pulse Oximeter
<b>Origin</b>	User

<b>Variable</b>	<b>Social Media Usage</b>
<b>Service</b>	RescueTime
<b>Description</b>	Your social media usage as measured by rescuetime
<b>Origin</b>	System

<b>Variable</b>	<b>Computer Productivity</b>
<b>Service</b>	RescueTime
<b>Description</b>	The total number of productive hours spent online as measured by Rescue time. Productive hours are those activities marked with productivity values of 1 or 2
<b>Origin</b>	System

<b>Variable</b>	<b>How often is Chloe vomitting?</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	User

<b>Variable</b>	<b>How many BM did Chloe have today?</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	User

<b>Variable</b>	<b>How much sleep did Chloe get at night?</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	User

<b>Variable</b>	<b>Deep Sleep</b>
<b>Service</b>	SMS
<b>Description</b>	Manually report your deep sleep hours from the Zeo device
<b>Origin</b>	User

<b>Variable</b>	<b>How long did Chloe nap?</b>
<b>Service</b>	SMS
<b>Description</b>	
<b>Origin</b>	User

<b>Variable</b>	<b>Inhalation Volume</b>
<b>Service</b>	SMS
<b>Description</b>	Inhalation volume measured in ml
<b>Origin</b>	User

<b>Variable</b>	<b>Appetite</b>
<b>Service</b>	SMS
<b>Description</b>	Measure of appetite/willingness to eat 5 - Shows signs of hunger, asks for food, eats all or most of meals/snacks 4 - Eats when given food but not actively expressing hunger 3 - Eats some of meals or snacks but shows reluctance 2 - Very little desire to eat but will drink Ensure or have an occasional snack 1 - No interest/desire to eat or drink anything but juice
<b>Origin</b>	User

<b>Variable</b>	<b>Sleep Quality</b>
<b>Service</b>	SMS
<b>Description</b>	A subjective measure of the quality of the night's sleep. Report how you feel after your night's sleep. It's ok if you had good quality, but not enough sleep and thus report feeling more tired. Your numbers need only be consistent for you - the absolute values don't matter greatly. You can use the journal page to provide a more detailed report.
<b>Origin</b>	User

<b>Variable</b>	<b>Energy Level</b>
<b>Service</b>	SMS
<b>Description</b>	On a scale of 1-5 where one is totally dragging and 5 is full of energy, how are you feeling?
<b>Origin</b>	User

<b>Variable</b>	<b>Cough Frequency</b>
<b>Service</b>	SMS
<b>Description</b>	With a baseline of no coughing, on a 7 point scale what is the coughing frequency in a given day with 0 being no cough and 7 being near constant.
<b>Origin</b>	User





# Bibliography

- [ACBF02] P Auer, N Cesa-Bianchi, and P Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2):235–256, 2002.
- [Ano09] UMLS Metathesaurus. U.S. National Library of Medicine, 2009.
- [Bac90] F Bacon. *Novum Organon* [1620]. *The Works*, 1990.
- [Ber10] Jesse A Berlin. N-of-1 clinical trials should be incorporated into clinical practice. *Journal of Clinical Epidemiology*, 63(12):1283–1284, December 2010.
- [BLH10] J Brubaker, C Lustig, and G Hayes. PatientsLikeMe: Empowerment and Representation in a Patient-Centered Social Network. In *CSCW’10; Workshop on Research in Healthcare: Past, Present, and Future*, Savannah, GA, February 2010.
- [BNJ03] D Blei, A Ng, and M Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 2003.
- [BNW08] D M Berwick, T W Nolan, and J Whittington. The Triple Aim: Care, Health, And Cost. *Health Affairs*, 27(3):759–769, May 2008.
- [Bro68] B Brown. DELPHI PROCESS: A Methodology Used for the Elicitation of Opinions of Experts. Technical Report RAND-P-3925, Santa Monica, CA, 1968.
- [BSK<sup>+</sup>] D A Barker, C C Sigman, G J Kelloff, N M Hylton, D A Berry, and L J Esserman. I-SPY 2 Trial. *Clinical pharmacology and therapeutics*, 86(1):97–100.
- [CCM04] W Cohen, V Carvalho, and T Mitchell. Learning to classify email into “speech acts”. *Proceedings of EMNLP*, 2004.
- [CHS<sup>+</sup>12] Connie Chen, David Haddad, Joshua Selsky, Julia E Hoffman, Richard L Kravitz, Deborah E Estrin, and Ida Sim. Making sense of mobile health data: an open architecture to improve individual- and population-level health. *Journal of Medical Internet Research*, 14(4):e112, 2012.

- [CK03] Jane Choi and John Y M Koo. Quality of life issues in psoriasis. *Journal of the American Academy of Dermatology*, 49(2):57–61, August 2003.
- [CNP02] N Chang, S Narayanan, and M Petruck. Putting Frames in Perspective. In *Proceedings of the 19th international conference on computational linguistics*, pages 1–7, 2002.
- [Coh88] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
- [Con03] Steve Connor. Glaxo chief: Our drugs do not work on most patients. *The Independent*, December 2003.
- [Cor] Matthew Cornell. Edison: The Experimenters Journal.
- [CW87] Charles W Champ and William H Woodall. Exact results for shewhart control charts with supplementary runs rules. *Technometrics*, 29(4), November 1987.
- [CWL<sup>+</sup>08] G Cong, L Wang, C Lin, Y Song, and Y Sun. Finding question-answer pairs from online forums. *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, pages 467–474, 2008.
- [CYR<sup>+</sup>07] David Cella, Susan Yount, Nan Rothrock, Richard Gershon, Karon Cook, Bryce Reeve, Deborah Ader, James F Fries, Bonnie Bruce, and Mattias Rose. The Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(Suppl 1):S3–S11, May 2007.
- [DAA<sup>+</sup>08] K Dwan, D Altman, J Arnaiz, J Bloom, and A Chan. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, 2008.
- [DEG<sup>+</sup>13] Naihua Duan, Ian Eslick, Nicole B Gabler, Heather C Kaplan, Richard L Kravitz, Eric B Larson, Wilson D Pace, Christopher H Schmid, Ida Sim, and Sunita Vohra. *Design and Implementation of N-of-1 Trials: A User’s Guide*. Agency for Healthcare Research and Quality (US), Rockville, MD, October 2013.
- [Des99] Rene Descartes. Discourse on Method and Related Writings. *Penguin Classics*, 1999.
- [DJA93] N Dahlbäck, A Jönsson, and L Ahrenberg. Wizard of Oz studies — why and how. *Knowledge-Based Systems*, 6(4):258–266, December 1993.
- [ECDK04] O Etzioni, M Cafarella, D Downey, and S Kok. Web-scale information extraction in knowitall:(preliminary results). *Proceedings of the 13th . . .*, 2004.
- [ECL11] Ian Eslick, CCHMC, and Lybba. Personal Experiments, October 2011.

- [EFT<sup>+</sup>12] Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, Adrian Edwards, and Michael Barry. Shared decision making: a model for clinical practice. *Journal of General Internal Medicine*, 27(10):1361–1367, October 2012.
- [EL05] I Eslick and H Liu. Langutils: A natural language toolkit for common lisp. In *Proceedings of the International Conference on Lisp*, 2005.
- [ELC12] Ian Eslick, Lybba, and CCHMC. MyIBD, September 2012.
- [ES10] D Estrin and I Sim. Open mHealth architecture: an engine for health care innovation. *Science(Washington)*, 330(6005):759–760, 2010.
- [Eys04] G Eysenbach. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, 328(7449):1166–0, May 2004.
- [FD13a] Susannah Fox and Maeve Duggan. Health Online 2013. *Pew Internet and American Life Project*, January 2013.
- [FD13b] Susannah Fox and Maeve Duggan. Tracking for Health. *Pew Internet and American Life Project*, January 2013.
- [Fel05] S R Feldman. Psoriasis assessment tools in clinical trials. *Annals of the Rheumatic Diseases*, 64(Supplement II):ii65–ii68, March 2005.
- [Fie06] Stephen E Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1–40, March 2006.
- [FKG<sup>+</sup>13] Ruth R Faden, Nancy E Kass, Steven N Goodman, Peter Pronovost, Sean Tunis, and Tom L Beauchamp. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *The Hastings Center report*, 2013(Jan-Feb):S16–27, January 2013.
- [FMWH08] Jeana H Frost, Michael P Massagli, Paul Wicks, and James Heywood. How the Social Web Supports patient experimentation with a new therapy: The demand for patient-controlled and patient-centered informatics. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, pages 217–221, 2008.
- [FOV<sup>+</sup>11] Jeana Frost, Sally Okun, Timothy Vaughan, James Heywood, and Paul Wicks. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. *Journal of Medical Internet Research*, 13(1):e6, 2011.
- [Fox09] S Fox. The social life of health information. *Pew Internet and American Life Project*, 2009.

- [GCB<sup>+</sup>03] Alice B Gottlieb, Umesh Chaudhari, Daniel G Baker, Michelle Perate, and Lisa T Dooley. The National Psoriasis Foundation Psoriasis Score (NPF-PS) system versus the Psoriasis Area Severity Index (PASI) and Physician's Global Assessment (PGA): a comparison. *Journal of drugs in dermatology : JDD*, 2(3):260–266, June 2003.
- [GDL<sup>+</sup>09] N B Gabler, Naihua Duan, D Liao, J G Elmore, and T G Ganiats. Dealing with heterogeneity of treatment effects: is the literature up to the challenge. *Trials*, 10(1):43, 2009.
- [GGK07] G Gigerenzer, W Gaissmaier, and E Kurz. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.
- [Gly03] C Glymour. Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, 7(1):43–48, 2003.
- [GM02] R Girju and D Moldovan. Mining Answers for Causation Questions. *AAAI Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
- [Gol12] Isaac Golden. Evidence-based research in complementary and alternative medicine I: history. *Journal of Evidence-based Complementary and Alternative Medicine*, 17(1):72–75, 2012.
- [GS86] B Grosz and C Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
- [GSA<sup>+</sup>88] G Guyatt, D Sackett, J Adachi, R Roberts, J Chong, D Rosenbloom, and J Keller. A clinician's guide for conducting randomized trials in individual patients. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 139(6):497–503, September 1988.
- [GSBT04] Griffiths, M Steyvers, D Blei, and J Tenenbaum. Integrating topics and syntax. *Advances in neural information processing systems*, pages 537–544, 2004.
- [GTL<sup>+</sup>07] Joel M Gelfand, Andrea B Troxel, James D Lewis, Shanu Kohli Kurd, Daniel B Shin, Xingmei Wang, David J Margolis, and Brian L Strom. The risk of mortality in patients with psoriasis: results from a population-based study. *Archives of dermatology*, 143(12):1493–1499, December 2007.
- [Hin08] Patricia H Hinchey. *Action Research*. Primer. Peter Lang, 2008.
- [Ins01] Institute of Medicine (US). Committee on Quality of Health Care in America. *Crossing the Quality Chasm*. 2001.
- [Ioa05] John P A Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, August 2005.

- [Ioa08] J Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648, 2008.
- [IR] Inspire, Inc and Manhattan Research. Lack of support drives social media use among psoriasis patients.
- [Iso09] Isometrik. The female menstrual cycle. Wikipedia, the Free Encyclopedia, December 2009.
- [JIB07] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, March 2007.
- [JNRM99] R Jones, K Nigam, E Riloff, and A McCallum. Bootstrapping for Text Learning Tasks. In *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, 1999.
- [Jon05] R Jones. *Learning to extract entities from labeled and unlabeled text*. PhD thesis, Carnegie Mellon University, May 2005.
- [JRGLL08] Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern Epidemiology*. 2008.
- [Jur02] D Jurafsky. Pragmatics and Computational Linguistics. *Handbook of Pragmatics*. Blackwell, 2002.
- [KBK08] A Keselman, A C Browne, and D R Kaufman. Consumer Health Information Seeking as Hypothesis Testing. *Journal of the American Medical Informatics Association*, 15(4):484–495, April 2008.
- [KCF<sup>+</sup>06] J Kim, G Chern, D Feng, E Shaw, and E Hovy. Mining and assessing discussions on the web through speech act analysis. *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*, 2006.
- [KDK<sup>+</sup>] Richard L Kravitz, Naihua N Duan, Heather Kaplan, Ian Eslick, Ida Sim, E B Larson, Sunita Vohra, William C Schmidt, and N B Gabler. DEcIDE Single Patient Trial Monograph. Technical Report TBD, Washington DC.
- [KF09] D Koller and N Friedman. *Probabilistic graphical models: Principles and techniques*. MIT Press (MA), 1st edition, 2009.
- [KM06] V Krishnan and C Manning. An effective two-stage model for exploiting non-local dependencies in named entity . . . . *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006.

- [KMB<sup>+</sup>05] Boris Katz, Gregory Marton, Gary Borchardt, Alexis Brownell, Sue Felshin, Daniel Loreto, Jesse Louis-Rosenberg, Ben Lu, Federico Mora, Stephan Stiller, Ozlem Uzuner, and Angela Wilcox. External knowledge sources for question answering. *TREC*, 2005.
- [Kol99] D Koller. Probabilistic relational models. *Inductive Logic Programming, Lecture Notes in Computer Science*, 1634:3–13, 1999.
- [Kru13] John K Kruschke. Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General*, 142(2):573–603, 2013.
- [KS86] R Kowalski and M Sergot. A logic-based calculus of events. *New generation computing*, 4:67–95, 1986.
- [KSDK08] A Keselman, C Smith, G Divita, and H Kim. Consumer Health Concepts That Do Not Map to the UMLS: Where Do They Fit? *Journal of the American Medical Informatics Association*, 15(4):496–505, 2008.
- [LAM09] LAM Foundation. LAM Foundation Listserv. LAM Foundation, 2009.
- [Lar93] Eric B Larson. Randomized Clinical Trials in Single Patients During a 2-Year Period. *Journal of the American Medical Association*, 270(22):2708, December 1993.
- [LC07] Annie Y S Lau and Enrico W Coiera. Do people experience cognitive biases while searching for information? *Journal of the American Medical Informatics Association : JAMIA*, 14(5):599–608, September 2007.
- [LLS11] Jingjing Liu, Alice Li, and Stephanie Seneff. Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs. In *IMMM 2011*, pages 91–96. IMMM 2011 : The First International Conference on Advances in Information Mining and Management, 2011.
- [LMP01] J Lafferty, A McCallum, and F Pereira. Conditional random fields. *Proceedings of the 18th international conference on machine learning*, pages 282–289, 2001.
- [LPD<sup>+</sup>11] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized Medicine*, 8(2):161–173, March 2011.
- [LPFH04] J Lester, S Prady, Y Finegan, and D Hoch. Learning from e-patients at Massachusetts General Hospital. *British Medical Journal*, 328(7449):1188–1190, May 2004.
- [Mar08] G Martin. ”Ordinary people only”: knowledge, representativeness and the publics of public participation in healthcare. *Sociology of health and illness*, 30(1):35–54, January 2008.

- [McC02] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. 2002.
- [MCM<sup>+</sup>12] Anmol Madan, Manuel Cebrian, Sai Moturu, Katayoun Farrahi, and Alex Sandy Pentland. Sensing the Health State of a Community. *IEEE Pervasive Computing*, 11(4):36–45, 2012.
- [Min07] M Minsky. *The Emotion Machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007.
- [ML09] John Moore and Henry Lieberman. Talking about painful subjects: flexibility and constraints in patient interviews. *Studies in health technology and informatics*, 149:130–139, 2009.
- [MPS13] P A Margolis, L E Peterson, and M Seid. Collaborative Chronic Care Networks (C3Ns) to Transform Chronic Illness Care. *Pediatrics*, 131(Supplement):S219–S223, May 2013.
- [MS04] A McCallum and C Sutton. Collective Segmentation and Labeling of Distant Entities in Information Extraction. Technical Report TR-04-49, 2004.
- [Mue06] E Mueller. Event calculus and temporal action logics compared. *Artificial Intelligence*, 170(11):1017–1029, 2006.
- [Nam11] Priya Nambisan. Information seeking and social support in online health communities: impact on patients’ perceived empathy. *Journal of the American Medical Informatics Association*, 18(3):298–304, May 2011.
- [Nat12] National Psoriasis Foundation. Psoriasis and Comorbid Conditions Issue Brief, January 2012.
- [NEC<sup>+</sup>10] Michael Nurok, Ian Eslick, Carlos R R Carvalho, Ulrich Costabel, Jeanine D’Armiento, Allan R Glanville, Sergio Harari, Elizabeth P Henske, Yoshikazu Inoue, Simon R Johnson, Jacques Lacronique, Romain Lazor, Joel Moss, Stephen J Ruoss, Jay H Ryu, Kuniaki Seyama, Henrik Watz, Kai-Feng Xu, Elizabeth L Hohmann, and Frank Moss. The International LAM Registry: a component of an innovative web-based clinician, researcher, and patient-driven rare disease research platform. *Lymphatic research and biology*, 8(1):81–87, March 2010.
- [NMS<sup>+</sup>10] Jane Nikles, Geoffrey K Mitchell, Philip Schluter, Phillip Good, Janet Hardy, Debra Rowett, Tania Shelby-James, Sunita Vohra, and David Currow. Aggregating single patient (n-of-1) trials in populations where recruitment and retention was difficult: The case of palliative care. *Journal of Clinical Epidemiology*, pages 1–10, October 2010.
- [Pag08] J Pagano. *Healing Psoriasis: The Natural Alternative*. John Wiley and Sons, 2008.

- [Pan98] R Panko. What we know about spreadsheet errors. *Journal of End User Computing*, 10(2):15–21, 1998.
- [Par11] Ashley Parker. Twitter’s Secret Handshake. *The New York Times*, June 2011.
- [Pea00] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- [PS88] J Pearl and G Shafer. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufman, 1988.
- [PW10] Michael Massagli Jeana Frost Catherine Brownstein Sally Okun Timothy Vaughan Richard Bradley James Heywood Paul Wicks. Sharing Health Data for Better Outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2), 2010.
- [RACB11] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. Passive and In-Situ assessment of mental and physical well-being using mobile sensors. In *UbiComp ’11: Proceedings of the 13th international conference on Ubiquitous computing*. ACM Request Permissions, September 2011.
- [RD06] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.
- [Rep09] Consumer Reports. What’s behind our dietary supplements coverage, January 2009.
- [Rob10] S Roberts. The unreasonable effectiveness of my self-experimentation. *Medical Hypotheses*, 75(6):482–489, 2010.
- [Rob12] Seth Robertson. Experimentation in the quantified self. September 2012.
- [Rot05] P M Rothwell. External Validity of randomised controlled trials:“To whom do the results of this trial apply?”. *Lancet*, 365(9453):82–93, 2005.
- [SAR<sup>+</sup>07] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.
- [SG07] M Steyvers and T Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.
- [SGS00] P Spirtes, C Glymour, and R Scheines. *Causation, Prediction, and Search*. The MIT Press, 2nd edition, 2000.



- [She39] W A Shewhart. *Statistical Method from the Viewpoint of Quality Control - Walter Andrew Shewhart - Google Books*. Graduate School of the Department of Agriculture, Washington DC, 1939.
- [SHSU06] T Sibanda, T He, P Szolovits, and O Uzuner. Syntactically-informed semantic category recognizer for discharge summaries. *AMIA Annual Symposium Proceedings 2006*, pages 714–718, 2006.
- [SL11] Dejun Su and Lifeng Li. Trends in the Use of Complementary and Alternative Medicine in the United States: 2002–2007. *Journal of Health Care for the Poor and Underserved*, 22(1):296–310, 2011.
- [SM05] L Shi and R Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Computational Linguistics and Intelligent Text Parsing*, pages 100–111. Springer Verlag, Berlin Heidelberg, 2005.
- [SRR05] L Slaughter, C Ruland, and A Rotegård. Mapping Cancer Patients’ Symptoms to UMLS Concepts. *AMIA Annual Symposium Proceedings 2005*, pages 699–703, 2005.
- [SSSM] Mark Smith, Robert Saunders, Leigh Stuckhardt, and Michael J McGinnis. *The Best Care at the Lowest Cost: The Path to Continuously Learning Health Care in America*. Institute of Medicine of the National Academies.
- [Sto00] Arthur A Stone. *The science of self-report: Implications for research and practice*. Lawrence Erlbaum Associates, Inc., 2000.
- [Sur05] James Surowiecki. *The Wisdom of Crowds*. Random House LLC, August 2005.
- [SW08] CA Smith and PJ Wicks. PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy. In *AMIA Annual Symposium Proceedings*, page 682, 2008.
- [Swa09] Melanie Swan. Emerging Patient-Driven Health Care Models. *International Journal of Environmental Research and Public Health*, 6(2):492–525, February 2009.
- [Swa12a] M Swan. Sensor Mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, 1(3):217–253, 2012.
- [Swa12b] Melanie Swan. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research*, 14(2):e46, 2012.
- [Swa12c] Melanie Swan. Scaling crowdsourced health studies: the emergence of a new form of contract research organization. *Personalized Medicine*, 9(2):223–234, March 2012.

- [TDSM06] Angelo M Taveira-DaSilva, Wendy K Steagall, and Joel Moss. Lymphangioliomyomatosis. *Cancer control : journal of the Moffitt Cancer Center*, 13(4):276–285, September 2006.
- [VdN07] W Van den Noortgate. The Aggregation of Single-Case Results Using Hierarchical Linear Models. *Behavior Analyst Today*, 8(2):52–57, 2007.
- [vH05] Eric von Hippel. *Democratizing Innovation*. MIT Press, 2005.
- [vH09] Eric von Hippel. Democratizing Innovation: The Evolving Phenomenon of User Innovation. *International Journal of Innovation Science*, 1(1):29–40, July 2009.
- [WGA07] Paul Wicks and A Gutierrez-Alvarez. Excessive yawning is common in the bulbar-onset form of ALS. Author’s reply. *Acta psychiatrica Scandinavica*, 116(1):77–78, 2007.
- [WM04] S Wan and K McKeown. Generating overview summaries of ongoing email thread discussions. *COLING’04: Proceedings of the 20th international conference on Computational Linguistics*, 2004.
- [WM05] X Wang and A McCallum. A note on topical n-grams. Technical Report UM-CS-2005-071, Massachusetts University Amherst Department of Computer Science, 2005.
- [WM11] D Walkiewicz and P Margolis. ImproveCareNow. *Inflammatory Bowel Disease*, 2011(17):450–457, 2011.
- [WMKD11] Paul Wicks, Michael Massagli, Amit Kulkarni, and Homa Dastani. Use of an online community to develop patient-reported outcome instruments: the Multiple Sclerosis Treatment Adherence Questionnaire (MS-TAQ). *Journal of Medical Internet Research*, 13(1):e12, 2011.
- [Wol10] Gary Wolf. The Data-Driven Life. *The New York Times Magazine*, May(2), May 2010.
- [Wol11] Gary Wolf. Experimental methodology in the quantified self. January 2011.
- [WSW08] J Weissman, E Schneider, and S Weingart. Comparing patient-reported hospital adverse events with medical record review: Do patients know something that hospitals do not? *Annals of Internal Medicine*, 149(2):100–108, 2008.
- [WVM11] Paul Wicks, T Vaughan, and M Massagli. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, 29(5):411–414, 2011.
- [YYP08] M Yetisgen-Yildiz and W Pratt. Finding the Meaning of Medical Concept Correlations. *AMIA Annual Symposium Proceedings*, 2008:830–834, 2008.

- [Zie03] R Zielstorff. Controlled Vocabularies for Consumer Health. *Journal of biomedical informatics*, 36(4):326–333, 2003.
- [ZRB03] Xiangjian Zheng, Sagarika Ray, and Wendy B Bollag. Modulation of phospholipase D-mediated phosphatidylglycerol formation by differentiating agents in primary mouse epidermal keratinocytes. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1643(1-3):25–36, December 2003.
- [ZRS10] Deborah R Zucker, Robin Ruthazer, and Christopher H Schmid. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of Clinical Epidemiology*, 63(12):1312–1323, December 2010.
- [ZSM<sup>+</sup>97] D R Zucker, C H Schmid, M W McIntosh, R B D’Agostino, H P Selker, and J Lau. Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology*, 50(4):401–410, April 1997.