

Personalized health experiments to optimize well-being and enable scientific discovery

by
Ian Scott Eslick

Thesis proposal submitted to the Program in Media, Art and Sciences,
School of Architecture and Planning, 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, high-quality clinical research, so patient's anecdotal experiences provide the only evidence. No method exists to help patients use this evidence to make decisions about their own care. My research aims to bridge the massive gap between clinical research and anecdotal evidence by putting the tools of science into the hands of patients.

Specifically, I will enable patient communities to convert anecdotes into structured self-experiments that apply to their daily lives. A self-experiment, a sub-type of the single-subject (N-of-1) trial, can quantify the effectiveness of a lifestyle intervention on one patient. The patient's challenge is deciding which of many possible experiments to try given the information available. A recommender system will aggregate experimental outcomes and background information from many patients to recommend experiments for each individual. Unusual interventions that succeed over many trials become evidence to motivate future clinical research.

Advisers and Readers

Adviser: _____

Frank Moss
Professor of the Practice of Media Arts and Sciences
MIT Media Laboratory

Reader: _____

Peter Szolovits
Professor of Computer Science and Engineering
MIT Department of EECS

Reader: _____

Henry Lieberman
Principle Research Scientist
MIT Media Laboratory

Table Of Contents

1	Summary	3
1.1	Self-Experiments and the Single-Subject Trial	3
1.2	Aggregating Self-Experiments	5
1.3	Hypothesis	5
2	Approach	6
2.1	Representing Self-Experiments	6
2.2	Recommending Experiments	7
2.3	The Cold Start Problem	8
3	Example	8
3.1	A Small Population Study	9
3.2	As a Self-Experiment	9
3.3	What to try first?	11
4	Prior Work	12
4.1	LAMsight: Community Data Collection	12
4.2	Patient Forum Data Mininig	13
5	Implementation	14
5.1	Tool Evaluation	14
5.2	Evaluating Recommendations	14
5.3	Tools	15
5.4	Schedule	16
6	Anticipated Outcome	16
6.1	Relevance	17
7	Biographies	20

1 Summary

The behavior and decision-making of healthcare providers are primarily driven by the outcomes of randomized controlled trials (RCTs) 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 can't provide a return on investment are unlikely to be run, leaving many important questions unasked.

When clinical science has nothing to say, patient anecdotes provide an alternate source of hypotheses; people are turning to this source of information to guide their personal health decisions in increasing numbers [Fox09]. This trend represents both opportunity and challenge. Patient-reported health information suffer from a multitude of biases and confounding factors [Sto00]. Even true outcomes do not generalize what works for one patient may not work for another. Yet, millions of people use anecdotal information coupled with guesswork to self-diagnose, self-treat and self-advocate.

Online anecdotal information takes multiple forms. Most users share information through natural language posts, such as opinions about disease phenomenology (“a gluten free diet will reduce symptoms of psoriasis”) or narratives about their own outcomes (“I tried a gluten free diet and my psoriasis got better”). A small, but growing, minority record quantitative data drawn from their everyday lives.

A collection of emerging, web-based platforms aggregate patient health data such as symptoms and medication use. [EA11] [Pat11]. Other sites elicit direct opinions from patient populations to accumulate ratings about treatment efficacy [Cur11]. A few sites are starting to propose novel methods for running population trials online [WVM11] [Gen11a]. All of these sites focus on serving the patient through visualizing and sharing population 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 evaluation and decision making entirely to the individual.

One solution to these problems involves re-framing patient self-report as the outcome of one or more self-experiments. A self-experiment quantifies anecdotes and defines the conditions under which data was collected, removing many sources of variation. A self-experiment is also proscriptive; running it tells the patient whether a treatment positively influences their symptoms or not. A patient can prove, to themselves and others, that their anecdotes are valid observations.

1.1 Self-Experiments and the Single-Subject Trial

The self-experiment is a subtype of the single-subject (“N-of-1”) trial. A single-subject trial uses the patient as their own control, comparing a past state to a future state that is the result of some controlled intervention. Generally, a trial represents:

- A hypothesis that a given treatment will influence an outcome;
- A protocol for testing the hypothesis including selection of measurement instruments, an intervention

schedule, and dosing; and

- An analytical method for evaluating effectiveness; “did this treatment significantly improve my outcome?”

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 [Hos11]. However, single-subject designs are only applicable when certain conditions hold [GSA⁺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.

Self-experiments lack the controls of single-subject trial; 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 effect¹.

¹Interestingly, the self-experiments in my research won't suffer from the under-reporting bias that regularly occurs in peer-reviewed literature [DAA⁺08] [Ioa08]. For some problems a series of self-experiments may outperform clinical research in both efficiency and quality

1.2 Aggregating Self-Experiments

In traditional clinical science, meta-analysis aggregates multiple population trials to characterize unexpected variation among similar trials or to improve estimates of the true 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⁺97] [dN07] [NMS⁺10].

The system I propose succeeds when a patient runs an experiment that demonstrates improvement in the chosen outcome. The burden of validity rests on the experiment itself. The patient decides which of many possible experiments to try, runs them, and repeats until a successful trial occurs. At one extreme the user randomly chooses experiments. Supporting the patient’s decision is equivalent to minimizing the number of trials needed to find a successful experiment. This is a weaker aim than traditional meta-analysis and an example of a “soft fail” system, one that is robust to noise.

The number of failed experiments tried may be reduced through discussion with other patients or using the outcomes of other patient’s experiments to predict a given patient’s response. This is the central question: how do we aggregate self-experimental results to best inform a patient’s decision?

Ideally, the patient would choose the treatment and experiment that maximizes the expected value of the patient’s response to treatment, given all available evidence. There are many sources of uncertainty that make direct calculation of these expectations extremely challenging in practice (e.g. use of multiple outcome instruments). Moreover, the system should account for likely adherence to the experimental protocol, a user’s preferences for treatments (e.g. medications vs. lifestyle), and minimize potentially undesirable side-effects. Developing an explicit model to simultaneously compute all these factors across dozens of treatments is likely to be intractable. At best, we can compute a rough approximation to the true distributions.

There are many ways to compute this estimate, such as exploiting existing techniques for meta-analysis (e.g. heirarchical linear or Bayesian models). Responses of similar patients are another strategy for computing likely response. For example, the experimental outcomes of a twin with exactly the same background as my own would be a good approximation. Other heuristic methods, such as global experimental success frequencies, may be used.

A recommender system perfectly fits the requirements of a decision support model for experiments. Recommender systems are typically ensembles of models, trained to fuse a diverse source of partially correct signals into a single system. These systems leverage many of the same techniques we encounter in epidemiology and clinical science (classification, similarity measures, hierarchical linear regression, Bayesian models, etc), but are designed to scale to millions of users by hundreds of thousands of items and be continuously improved through user action over time.

1.3 Hypothesis

Consequently, the thesis will evaluate the following assertions:

- A large fraction of patient anecdotes can be practically translated or annotated by patients to create structured self-experiments that form a space of treatment hypotheses;

- A patient can successfully execute these self-experiments to gain confidence in the effectiveness or ineffectiveness of a specific hypothesis; and
- Individual experimental outcomes can be aggregated in a recommender model that simplifies and optimizes an individuals search through the space of hypotheses.

An iterative model of recommendation plus self-experimentation will address problems where professional medical advice is not easily available or not easily applied, where self-experiments are an ethical solution to an individual problem, and where they can be practically applied. Ideally, the methodology outlined here will converge with efforts to make N:1 trials an extension of standard clinical practice.

2 Approach

To test my hypotheses, I will implement an online tool that enables motivated individuals to translate anecdotes into self-experiments. I will develop a representation of self-experiments that captures the range of symptoms and proposed treatment found in two online communities (Psoriasis and Crohns Disease) and work with real-world user communities to evaluate their effectiveness.

Self-experimental outcomes from these experiments will be aggregated using a recommender engine to provide early personalized patient recommendations. Long-term behavior of a community will be approximated through a simulation model that characterizes the sensitivity of the recommendation model to different impairments including: systematic reporting bias, false negatives and positives, and high instrument diversity.

Finally, I will address the practical “cold start” issues which plague recommender systems by exploiting existing information sources as seed evidence. The structured representation of self-experiments and a recommender system that exploits it are the main technical contributions of the thesis.

2.1 Representing Self-Experiments

The representation of self-experiments is a critical component of the thesis. A representation must support the wide range of possible experiments that user may want to run, yet avoid creating problems of sparsity (e.g. every user runs a unique experiment). Any single disease can have dozens of different treatments, each of which may be tested in different ways. For example, Table 1 lists the many treatments 200 patients on CureTogether.org collectively curated for Psoriasis.

The following list contains a high-level description of the elements that are important to running successful self-experiments and must be part of any representation of experiments:

1. **Outcome Variable** What measurable factor do we want to influence? Outcome variables include high level factors like sleep, fatigue, or other physical or mental symptoms;
2. **Intervention** What life factor will we manipulate to achieve a change in the outcome variable? Interventions include habit changes such as an earlier sleep time, dietary changes such as a gluten-free diet

Topical corticosteroids	Calcipotriene (Dovonex)	Psoriatec	Tazarotene (Tazorac, Avage)
Tacrolimus	Pimecrolimus	Cool tar	Moisturizer
Sunlight	UVA phototherapy	UVB phototherapy	Excimer laser
Oatmeal baths	Avoid alcohol	Avoid triggers	Jarrows probiotics
Greens+	Vitamin D	Abrasion	Urea-based moisturizer
Grapefruit seed extract	Methotrexate	Azathioprine	Etretinate
Hydroxyurea	Mycophenolate (Mofetil)	Cyclosporine	Healing Tree Psoriasis Ointment
Banana peel rub	Ustekinumab (Stelara)	Tegrin	Turmeric
Salicylic acid	Etanercept (Enbrel)	Aloe vera	Oil of Olay-sensitive skin soap and cream
Infliximab (Remicade)	Low Dose Naltrexone (LDN)	Regividerm	Calcipotriol+Betamethason
T-Gel	Anthralin (Dritho-Scalp)	Calcineurin inhibitors	Narrowband UVB therapy
Psoralen	Acitretin (Soriatane)	Alefacept (Amevive)	Daily baths
Epsom salts	Avoid hot water and harsh soaps	Avoid stress	Avoid smoking
Avoid sunburn	Avoid skin injury	Fish oil	25% Coal Tar
Loma Lux Psoriasis Homeopathic			

Table 1: Crowdsourced Taxonomy of Psoriasis Treatments

or increased fat intake, and supplementation such as St. John’s Wort. An intervention may involve discussing the experiment with your physician to try a prescription medication;

3. **Dose** When appropriate, the dosage and frequency must be specified.
4. **Confounders and Side-Effects (optional)** What other factors will we track to adjust for confounding in the outcome variable? What side-effects are potential problems or inconveniences should we track?
5. **Instruments** What methods will we use to measure outcomes, confounders, and side-effects? Instruments influence patient compliance as well as statistical significance, for example by introducing measurement noise. Instruments can be questionnaires, rating scales, or physical devices;
6. **Trial Type** How will we test our intervention? I plan to use simple reversal designs, such as ABA or ABAB-style designs analyzed by statistical tests or graphical inspection;
7. **Schedule (generated)** How often will we measure and how long are the control and intervention periods? Treatment and outcome factors (such as onset and offset times for treatment effect) determine the schedule.

The envisioned system supports a process of user collaboration; self-experiments are defined and debugged by user populations. Users bear responsibility for curating and improving experiments suffering from insufficient data, poor instruments, or other methodological problems. The collaborative model here is similar to wikipedia, and open discussion is encouraged about different elements of the self-experiment representation. Primary data will be recorded, eventually allowing sophisticated users to download and perform secondary analysis on the experimental data.

2.2 Recommending Experiments

A recommender system’s task is to reduce a large set of possible experiments to a small set that are interesting, practical and effective for a given individual. Each self-experiment outcome provides information relevant to all three factors. For example, non-compliance is a strong indication that the instruments used or

the required schedule are infeasible for that individual. The success of an experiment indicates both a useful methodology and an effective treatment. A curator may create a new experiment and declare that a prior positive result was a false-positive due to methodological error.

Collaborative filtering recommendation engines, such as those used by companies like Amazon [LS03], make extensive use of item-to-item similarities. The parameters of self-experiments, such as goals, treatments, instruments, and inheritance (i.e. where one experiment derives from another), enable combination of evidence from multiple experiments to guide recommendation. Executing multiple self-experiments with different parameter choices is analogous to a factorial design; the prediction system needs to identify the combination that is the most effective. Simple algorithms (such as [Lem05]) have been remarkably successful on very complex item-to-item comparison problems. Identifying the appropriate metrics of similarity for experimental templates is a key part of the research agenda.

2.3 The Cold Start Problem

Unfortunately, recommender systems typically require a significant number of users and user ratings to yield useful recommendations [Su. . . 09]. In the recommendation literature the large amount of data needed to produce useful results is called the “cold start” problem. To address the cold start problem I will leverage several strategies from the literature: introducing seed, or proxy, data, and exploiting experiment similarity.

Specifically, I believe that direct or derived assessments of belief in treatment efficacy will serve as the simplest and most valuable proxies for experimental outcomes. For example, CureTogether.org has engaged hundreds of patients who have evaluated the efficacy of dozens of treatments, including medication and lifestyle changes, across a hundred or more conditions. They also provide co-morbidity and background information.

Similarly, online patient forums are a sequence of questions, statements, case reports, and treatment suggestions that can be automatically converted into a per-user feature vector. User “Votes” over treatments will be identified via lexical-syntactic templates and looking at significant correlations of symptoms and treatments (see [Li11]).

These sources fall into a hierarchy of increasing specificity, but decreasing ease-of-acquisition,

- Existing online forums [Fou11] [oC11];
- Structured forums viewed as case reports [LLC11]
- Direct elicitation of beliefs [Cur11] [EA11]
- Longitudinal/observational records of treatment and outcomes [Pat11]
- Reports from self-experiments [Cor11]

These proxies, along with per-user background factors, give the recommender engine a seed dataset which boosts early performance, but becomes less relevant as experimental outcomes are accumulated.

3 Example

In mid-2010, psychologist and self-experimenter Seth Roberts gave a talk at a Quantified Self meetup. He found that eating 1/2 stick of butter each day improved his performance on a simple math test which he

had devised to evaluate his mental performance. He claimed a statistically significant decrease in average response time over several ABA-style self-experimental trials [Rob11].

The effect of butter fat on cognitive performance has not been studied in the clinical literature. Large population studies only recently began exploring the specific relationship between dietary fat and cognition, largely in the context of Alzheimers research. Fats influence cognition in many ways; it is plausible that consumption of specific fats have measurable influence on cognitive performance in other people, making this a suitable topic for self-experimentation.

3.1 A Small Population Study

In late 2010, Genomera [Gen11a], a startup focused on crowd-sourcing clinical research, independently sponsored a small population study [Gen11b] to see if a larger population demonstrated improved math scores after consumption of butter.

The study recruited 45 people and consisted of an ABA design with three treatment groups: a butter treatment group, a coconut oil treatment group and a control group. The participants were randomly assigned to treatment or control arms after the first week of baseline data was collected. The raw data and results of this study were made publicly available.

The study reported a mean 5% improvement in average response times in the butter group, a 2% improvement for the Coconut Oil group, and less than 1% improvement for the control group. Only the Butter group achieved significance ($p = 0.006$), but there was a high variance in the population between 1% and almost 10% improvement. A strong practice effect was observed; the final baseline phase was the best scoring of all three phases for nearly all participants.

This experiment required bringing a large number of people together at the same time, agreement on a single protocol, and adherence to the study terms. The result is a single analytical outcome that masked the high variation across the test subjects that emerges when analyzing individual data. Moreover, a follow up experiment has taken more than 3 months to assemble. What if each of these users had provided an independent test at different points in time without centralized control? What if a single experimenter had discovered the practice effect and modified the protocol for future participants? A set of related self-experimental trials may simplify data collection and compensate in part for the loss of randomization and other error-reducing benefits in centrally-managed population trials.

3.2 As a Self-Experiment

The Genomera study easily translates into two independent self-experiments as illustrated in Tables 2 and 3.

Figure 1 shows a Shewhart-style control-chart representation of one Butter experimenters from the Genomera study. It is a convenient way to see all the data and along with a significance test in a glance. Significance of an effect on a system in a control chart is determined when there are 2 or more points above or below a 3-sigma line from the historical mean. By this standard, the experiment failed to demonstrate significance.

Two sources of noise affected this chart and suggest modifications of the protocols. First, the dataset

Control Chart for fnmeyer

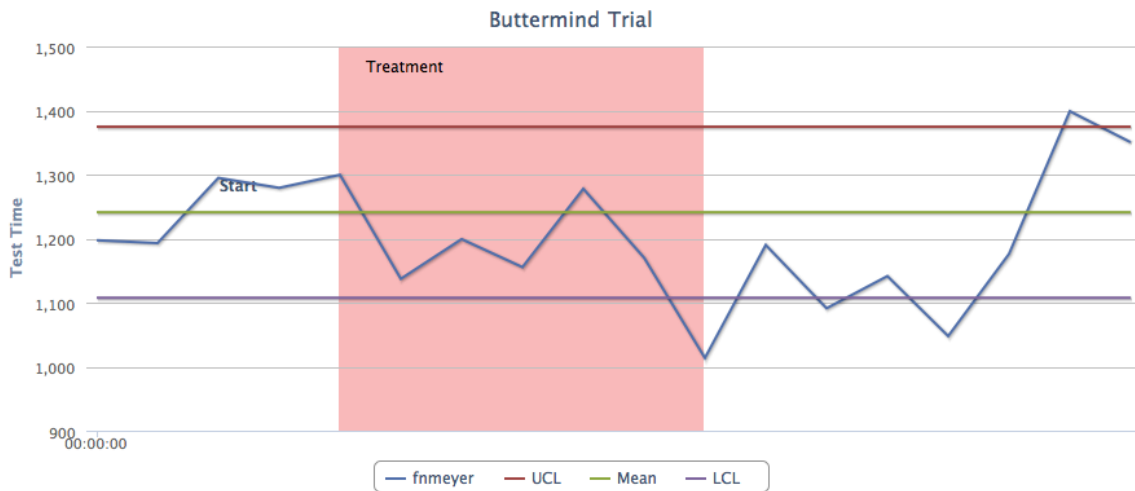


Figure 1: Single-user Control Chart

Goal	Improve mental performance
Variable	Mental performance
Instrument	Seth Robert's math test
Intervention	Eat Butter
Dose	4 Tbsp
Type	ABA
Schedule	1 week, 1 week, 1 week
Analysis	Control chart

Table 2: Butter Experiment

Goal	Improve mental performance
Variable	Mental performance
Instrument	Seth Robert's math test
Intervention	Eat Coconut Oil
Dose	4 Tbsp
Type	ABA
Schedule	1 week, 1 week, 1 week
Analysis	Control chart

Table 3: Coconut Experiment

demonstrates a strong practice effect. Removing the practice effect requires that math scores settle before starting an experiment. The documentation of the math-test instrument should record this requirement on behalf of all experiments it is used in, perhaps that 2 weeks of experience is required prior to starting the experiment.

Secondly, the analysis assumes the treatment has both zero onset and offset delay. The analysis can address this by shifting the treatment start window to overlap the expected start and end of the effect.

I expect that educated users who identify problems like the two above will create derivative self-experiments that fix problems as shown in Tables 4 and 5.

This small change illustrates the challenge of identifying a representation for self-experiments; a structured representation of a self-experiment must be flexible enough to a large fraction of likely experiments while being simple enough to use to aggregate outcomes from related self-experiments as discussed below.

The new self-experiment protocols shown in Tables 4 and 5 improve the originals and are more likely to produce true positive outcomes. Does this imply that we should throw away the 42 user outcomes from the Genomera experiment in preference of these new experiments? The Genomera study did show at the population level that butter is more likely to reduce test-taking time than coconut and in the absence of ad-

Goal	Improve mental performance
Variable	Mental performance
Instrument	Seth Robert’s math test
Intervention	Eat Butter
Dose	4 Tbsp
Type	ABA
Schedule	1 week, 1 week, 1 week
Analysis	Control chart, 2 day onset/offset
Pre-condition	2+ weeks experience with math test

Table 4: Butter Experimental (new)

Goal	Improve mental performance
Variable	Mental performance
Instrument	Seth Robert’s math test
Intervention	Eat Coconut Oil
Dose	4 Tbsp
Type	ABA
Schedule	1 week, 1 week, 1 week
Analysis	Control chart, 2 day onset/offset
Pre-condition	2+ weeks experience with math test

Table 5: Coconut Experiment (new)

	Butter	Coconut
Original	4/12	2/14
New	0/0	0/0

Table 6: Which experiment to recommend?

ditional information, a recommendation system should recommend Butter first - but which self-experiment? The older one with all the evidence, or the newer one? It is this question that the recommendation engine must answer.

3.3 What to try first?

A good self-experiment is relevant to a user’s goals and preferences, and likely to succeed when run. A recommendation algorithm must identify the associations among user features, experiment history, and item features that identifies good self-experiments. For example, Table 5 illustrates the available empirical evidence for the four self-experiments.

Old Butter may have the best empirical success rate, but according to users, New Butter is more likely to yield true positives/negatives. Because the treatment (Butter) is the same between the two experiments, it should have equivalent effectiveness to Old Butter at a similar sample size. Knowledge of the relationship between two experiments enables this inference. Generally, combining knowledge from similar experiments is critical to allowing a diversity of experiment types for any given treatment.

What if the user in question has a background factor, such as a specific disease or symptom, that is the same as the sub-population that showed a response to coconut oil? The recommendation system may prefer to suggest coconut oil as the first or second recommendation for this user. Similarly, if there is background information available, such as many people reporting online that coconut oil improved cognition, it can be used as a weak bias in the absence of more compelling outcome evidence.

This simple example illustrates aspects of the proposed work, specifically the dual constraint of structuring experiments to be useful, while enabling the different sources of information to be brought together into a single, predictive model.

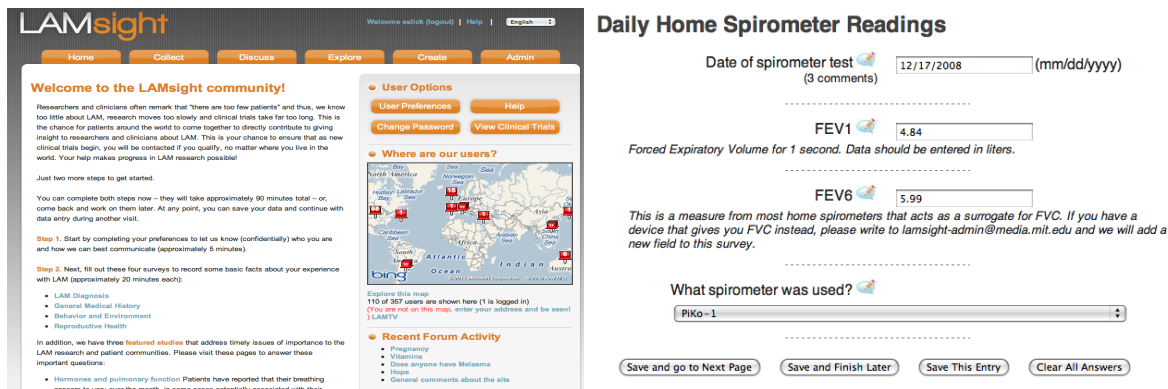


Figure 2: LAMsight home page and data journal

4 Prior Work

Over the past three years, I built and tested data collection systems similar to the one described above; I worked with patient populations to enable effective collaboration between patient communities and research professionals.

4.1 LAMsight: Community Data Collection

The LAM Treatment Alliance invests in research and infrastructure to accelerate the search for a treatment or cure for the rare disease Lymphangiomyomatosis (LAM). LAM is a complex, multi-system disorder that strikes women in their child-bearing years. It is characterized by the slow proliferation of smooth muscle tissue in the lungs, kidneys, and brain. LAM is debilitating. Decline is associated with a reduction in vascular activity, which reduces oxygen intake and significantly impairs activity. LAM is ultimately fatal through destruction of lung tissue. Lung transplantation is the only treatment for LAM, but LAM is metastatic and transplanted lungs invariably develop new LAM tissue. There are less than a thousand diagnosed patients with sporadic LAM in the United States today. (About 1/2 of women with Tuberos Sclerosis also develop LAM, but LAM tends to be of secondary concern to this population).

I developed a prototype web platform LAMsight (Figure 4.1) that:

- Tracks the worldwide population for this rare disease (word of mouth has made LAMsight the 3rd largest registry of LAM patients worldwide);
- Segment the population for trial recruitment and targeted questionnaires;
- Provide an open-ended survey creation mechanism to enable patient-directed collection of data;
- Support aggregation and exploration of population data by patients.

Patients and researchers have used the site to generate questionnaires that have uncovered and characterized novel observations about the disease pathology, particularly short-term variation in symptoms due

to the action of hormones. Over 100 patients have filled out extensive online surveys capturing data about sleep habits, rate of decline, and alternative medication use.

An ongoing experiment, “The LAM Estrogen Study” applies a novel subjective assessment instrument from another lung disease, COPD, to track the variation in dyspnea with the menstrual cycle. This was a common patient anecdote brought up at a conference I attended where an investigator suggested the method by which we might characterize the patient observation. A second patient-initiated investigation is exploring the effect of sexual activity on dyspnea.

LAM patients demonstrated tremendous interest in engaging with these questions, but burn-out was a significant factor and most patients failed to maintain long-term engagement. According to a survey of patients, they are conscious of their inability to sustain engagement day-to-day; however, all patients demonstrate a willingness and ability to engage semi-regularly when prompted by specific focused questions or activities. This led, in part, to the idea of allowing people to participate in short experiments at their own pace and accumulating evidence at a pace that works for the community, rather than on the investigator’s clock.

4.2 Patient Forum Data Mining

I built a framework to extract medical information from the LAM Foundation’s Listserv [oC11] to better understand the domain of patient anecdotes in LAM and other conditions. The LAM Listserv community consists of predominantly US-based individuals with LAM. The listserv archives are publicly available and a simple web scraper was built to extract message content. The corpus consists of 42,000 messages over the past 4 years. This comprises approximately 6 million total occurrences of 250k unique tokens.

I randomly sampled several hundred messages into speech acts, finding that 30% of forum content contained medically relevant information such as symptoms, treatments, or medical events. A manual phrase-level annotation identified syntactic contexts and lexical terms associated with disease and patient lifestyle knowledge.

I identified two major patterns of patient language use. Patients share generally accepted knowledge as factual statements, and their specific cases and explorations as narrative. I built a filter that selected for past tense language and narrative terms and identified a significant number of symptoms and treatments that were not part of the a standard medical database, yet frequently discussed by the community.

Finally, I experimented with automatic relation extraction. While I was regularly able to achieve a modest accuracy of 65-75% on my test sets, recall was barely above 10%. The remaining 90% of relationships between symptoms and treatments do have interesting distributional properties; some treatments are highly associated with certain symptoms (such as low-triglyceride diets for LAM patients with lymphatic leakage). The association of terms in different classes provide useful information (see [Li11]), but are only a rough guide to what treatments might be relevant. I realized that the weight of class association of a treatment can contribute to narrowing the field of focus for a self-experimenter.

The techniques just described yield similar results on other forums, such as cancer and psoriasis. The diseases may change, but the patient’s interest and language use appear remain extremely similar across domains.

5 Implementation

I plan to continue improving the existing web front-end to expose the tools and recommendation engine to a real-world population. For population experiments, I will seek collaborators with existing technologies (mobile apps, reminder systems, sensing devices, etc) to enable me to focus on the scientific rather than development aspects of this approach, but will build technology as necessary.

An initial experiment will expose a simple system with a set of self-experiments that I've manually derived from the psoriasis forums with initial recommendations derived from seed data. A SMS and web-based reminder and data capture system populates a database and provides simple individualized visualizations (control charts), and supports community discussion.

This research's biggest challenges are:

1. Finding a successful design point for the self-experiments
2. Demonstrating that, over time, a community of individuals running self-experiments will benefit from the experience of others.

5.1 Tool Evaluation

I will use a qualitative research approach to evaluate the self-experimental representation. Specifically, I am interested in the user's:

- Comprehension of the protocol;
- Compliance to the protocol;
- Assessment of the web platform;
- Analysis of relationships between user features, experimental outcomes, and stated preferences; and
- Explicit feedback on needed or unnecessary features.

I intend to run an ongoing observational experiment by exposing the tool to a community of real users to play with. I will also run specific, focused experiments on the platform to evaluate improvements to the template definitions and user interface or to characterize some aspect of user behavior.

5.2 Evaluating Recommendations

A recommender system is often evaluated through the use of a held-out dataset [HKB99]. The limited amount of data likely to be generated in early trials motivates a hybrid evaluation strategy.

- Subjective assessment. I will use existing voting data, such as that produced by CureTogether, or a survey of study participants to order experiments as a standard against which to compare to the recommendation's engine "cold start" state.

- Held out data. Sufficient, related experimental outcomes should enable me to perform a simple test of held out data on real world settings.
- Simulation. A generative user model will simulate user interactions and experimental responses to characterize the long-term behavior of the recommender system. I am most interested in quantifying the impact of bias, early false positives, and experiment proliferation (e.g. sparse data due to too many variations on a theme).
- Model validation. The user experiments described above will be used to identify the important parameters of a user model and to tune the simulation model.

The use of simulation to evaluate the long term behavior of social systems is a well-established practice [Gil00]. The approach suggested here is related; I will generate a set of 'experimenter' agents each of which has a parameterizable pattern of behavior. The behavior consists of two levels:

1. Physical disease model which dictates a response to an intervention, and
2. Impairments in measurement of that response (e.g. confounders that increase response variation).

Models are defined by a simple Bayesian network with random variables. The probability distributions of each instance of the network are chosen by meta parameters (e.g. sampled from a Beta distribution). An experiment in the simulation is the outcome of a sample of the disease and impairment models for each agent, followed by an update to the recommender system.

A sensitivity study models a sequence of self-experiments by creating an artificial population of agents; each agent engages in a search process, selecting individual experiments according to the recommender engine's recommendations and iterating until success or a pre-defined burn-out threshold is reached. Can the recommender system, as designed, discover the differences between populations of agents that respond differently to two treatments and reduce the average time to discovery?

Impairments shown by users in real experiments will motivate modifications to disease model, impairment model, and the self-experiment representation. Sensitivities exposed by the recommender system will motivate user studies or additional literature search to improve the characterization of how real users behave.

5.3 Tools

For the cold start problem and for community hypothesis generation, I've collected 30,000 forum messages from the leading US psoriasis forum to seed the recommender system. I have also gained access to Cure-Together's dataset of 150 psoriasis patient belief assessments regarding co-morbidity, treatments tried and their relative efficacy. There are other resources, such as EarthClinic.org, that provide collaboratively edited lists of treatment recommendations with up/down votes that can seed both the initial set of self-experiments and the recommendations.

The work discussed in section 4 provides a wide array of existing tools for analyzing new data. The existing tool suite consists of over a hundred thousand lines of Common Lisp code implementing web frameworks, databases, web mining tools, statistical and pattern-based NLP techniques, connections to external tools, and adapters for a variety of static datasets such as the UMLS, ConceptNet, WordNet, and standard statistical tools for computing basic properties of numeric datasets.

Support is being provided by a foundation, Lybba, and the Cincinnati Children's Hospital who are working on systematic implementation and scaling of clinical N-of-1 experiments in the home environment. Their evolving social platform provides an alternative setting for evaluating my research and may yield datasets for the recommender system as well.

Other partners such as members of Quantified Self, the ACOR cancer forums, LAM Treatment Alliance, and CureTogether.org have agreed to make data sets, time, and users available to support this research. Compass Labs, a small company which engages in large scale mining of social media and other web-based information sources, is making source code as well as physical infrastructure available to support this work.

5.4 Schedule

- April, 2011 - Submit proposal to department committee
- May, 2011 - Submit first experiments to COUHES
- May, 2011 - Thesis Proposal Defense
- June, 2011 - Recruit Psoriasis population from online forums
- July, 2011 - Run initial qualitative study to evaluate the use of experimental templates for Psoriasis patients
- July, 2011 - Finish development and initial runs of the simulation model of the recommender system
- August, 2011 - Release system into the wild for evaluation of natural uses of the system
- August/September, 2011 - Submit second qualitative study to evaluate the recommender system and needed changes to the template model
- September-November, 2011 - Run and analyze second user study and free-form use of the platform. Continue to develop representation and simulation model
- December, 2011 - Submit first draft of thesis to committee, identify open issues
- Spring, 2011 - Resolve open issues, run additional small studies as needed.
- June, 2012 - Thesis Completion and Defense

6 Anticipated Outcome

The application of a recommender system to aggregation of single patient experimental outcomes is a novel proposition. The significant contribution of the thesis is demonstrating that a sequence of self-experiments can successfully train a recommendation system to make accurate predictions of individual and global effectiveness of treatments, instruments, and protocols. The simulation model will show that, under the right conditions, the recommender system can converge to successfully:

1. Identify experiments relevant to user's symptoms;
2. Identify treatments relevant to a user's preferences;
3. Prioritize experiments when prognostic factors exist and are recorded;
4. Identify user preferences for different instruments or experiment properties; and

5. Tolerate the data sparsity observed in real-world experiments.

Establishing the real-world validity of the simulation requires deployment of a concrete system and extensive interaction with real users. The descriptive representation of self-experiments I develop will be an important secondary contribution.

Both the test platform and recommender engine prototypes will be made available as open source for others to build upon. I hope that one or more partner organizations will build on these ideas in parallel with my thesis timeline.

6.1 Relevance

Identifying methods to validly aggregate self-experiments represents a novel contribution to a world that is generating increasing amounts of data about people and concerned with behavior change.

Aggregating self-experiments is a novel and timely contribution. Evolution in technology is making it easier to capture and aggregate data, and rising healthcare costs demand more effective solutions for understanding and supporting behavior change that improves health outcomes.

Existing solutions have few proposals for determining treatment validity outside the clinical trial context. Models derived from aggregated self-experiments may prove more useful in clinical practice than the population-level models produced in most clinical trials. They represent a similar accumulation of experience to that doctors build over years of practice, but one easily available to everyone.

It is possible that the self-experimental model will remain an early-adopter phenomenon, limiting the direct applicability of my prototype. However, several large-scale trends suggest complementary means of acquiring similar information from the everyday experiences of individuals including:

- Sensor technologies and network technologies that lower the barriers to collecting quantifiable data about our everyday lives [LPAA09];
- Electronic traces of our online and daily lives reprocessed as valuable indicators of our health (blogs, social media postings, credit card transactions, etc);
- Open access to EMRs/PHRs that enable the integration of our medical history into prediction or recommendation models [KA05].

Pervasive, passive data collection facilities and lightweight reminder systems can make continuous self-improvement a part of the everyday experience and ideally, become a standard tool for traditional healthcare delivery.

Management of chronic health conditions is the focus of my proposal, but there are broader implications of this approach.

- **Optimizing well-being** Anyone wishing to understand and optimize their personal well-being will benefit from increased confidence that they are improving their condition;

- **Rationalizing alternative therapies** Alternative therapies are used by millions of people, yet few therapies have been or are likely to be the subject of clinical research. Having a large number of experimental outcomes developed independently of the vendor or practitioner can help an individual assess the general applicability of any given treatment.
- **Novel sources hypotheses** Surprising, yet reliable, self-experimental outcomes uncover holes in our understand of the human body and suggest new research directions;
- **New tools for scientific discovery** Interventions that work may be useful to the individual, but even a large set of successful self-experiments won't explain why they work. A population of people can run multiple experiments to test different hypotheses about the causal factor at work. For example, if I do math faster when I eat butter, do I see the same response with other saturated fats? With polyunsaturated fats? Is it Omega 3, 6 or their ratio?

References

- [Ber10] Jesse A Berlin. N-of-1 clinical trials should be incorporated into clinical practice. *Journal of Clinical Epidemiology*, 63(12):1283–1284, Dec 2010.
- [Cor11] Matthew Cornell. edison: The experimenter's journal. <http://edison.thinktrylearn.com/>, 2011 (accessed April, 2011).
- [Cur11] CureTogether.org. Curetogether home page. <http://curetogether.org/>, 2011 (accessed April, 2011).
- [DAA⁺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*, Jan 2008.
- [dN07] W Van den Noortgate. . . . The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, Jan 2007.
- [EA11] Ian Eslick and LAM Treatment Alliance. Lamsight. <http://www.lamsight.org>, 2007 (accessed April 19th, 2011).
- [Fou11] National Psoriasis Foundation. Talkpsoriasis.org. <http://www.inspire.com/groups/talk-psoriasis/>, 2011 (accessed April 19th, 2011).
- [Fox09] S Fox. . . . The social life of health information. *Pew Internet and American Life Project*, Jan 2009.
- [Gen11a] Genomera. Genomera.com. <http://genomera.com/about>, 2011 (accessed April 19th, 2011).
- [Gen11b] Eri Gentry. Butter mind conclusions/findings draft. <https://docs.google.com/document/d/1PePHcdLXQLZibINlePkLMcbynLLCSUS7qGwe7HXFY9g/edit?hl=en>, 2011 (accessed April 19th, 2011).
- [Gil00] N Gilbert. . . . How to build and use agent-based models in social science. *Mind & Society*, Jan 2000.
- [GSA⁺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*, 139(6):497–503, Sep 1988.
- [HKB99] J Herlocker, J Konstan, and A Borchers. . . . An algorithmic framework for performing collaborative filtering. *Proceedings of the 22nd . . .*, Jan 1999.

- [Hos11] Cincinnati Children’s Hospital. C3n. <http://c3nproject.org/>, 2011 (accessed April, 2011).
- [Ioa08] J Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, Jan 2008.
- [KA05] Isaac S Kohane and Russ B Altman. Health-information altruists—a potentially critical resource. *N Engl J Med*, 353(19):2074–7, Nov 2005.
- [Lem05] D Lemire. . . . Slope one predictors for online rating-based collaborative filtering. *Society for Industrial Mathematics*, Jan 2005.
- [Li11] Y Li. Medical data mining: Improving information accessibility using online patient drug reviews. *groups.csail.mit.edu*, Jan 2011.
- [LLC11] Earth Clinic LLC. Earthclinic.org. <http://earthclinic.org>, 2011 (accessed April 19th, 2011).
- [LPAA09] D Lazer, A Pentland, L Adamic, and S Aral. . . . Life in the network: the coming age of computational social science. *Science (New York)*, Jan 2009.
- [LS03] G Linden and B Smith. . . . Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing*, Jan 2003.
- [NMS⁺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, Oct 2010.
- [oC11] University of Cincinnati. Lam foundation listserv. <http://lam.uc.edu/html/listproc.html>, 2011 (accessed April 19th, 2011).
- [Pat11] Inc PatientsLikeMe. Patientslikeme. <http://patientslikeme.com>, 2011 (accessed April, 2011).
- [Rob11] Seth Roberts. Arithmetic and butter. <http://blog.sethroberts.net/2010/08/13/arithmetic-and-butter/>, 2010 (accessed April 25th, 2011).
- [Sto00] A Stone. The science of self-report: Implications for research and practice. *books.google.com*, Jan 2000.
- [Su. . . 09] X Su. . . . A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, Jan 2009.
- [WVM11] P WICKS, T Vaughan, and M Massagli. . . . Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, Jan 2011.
- [ZSM⁺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):401–10, Mar 1997.

7 Biographies

Ian Eslick



Ian Eslick earned his bachelor's and master's degrees from MIT's EECS department. He worked in MIT's Artificial Intelligence Laboratory as an undergraduate, then graduate, researcher focused on massively parallel computers, semiconductor devices, and enabling software. His M.Eng. thesis introduced a model for distributed, quasi-static compilation leveraging run-time program instrumentation.

In 1996, Ian co-founded Silicon Spice, a telecommunications semiconductor startup spun out of the AI Lab based in Mountain View, CA. As the founding President, Ian was responsible for 3 rounds of financing, hiring the company's management team, and played key roles in both product development and technical sales.

The 120 employee company was acquired by Broadcom Corporation in 2000 where, as Director of Software Engineering, he helped start Broadcom's Mobile Handset Business Unit.

Ian returned to MIT in 2003 as a Research Affiliate working with Patrick Winston at MIT's CSAIL. He later joined the Media Lab as a student in 2004 to work with Marvin Minsky and Push Singh in the Common Sense Computing project. His master's thesis developed semi-supervised techniques for mining large-scale semantic networks from the Web.

In 2006, Ian joined Frank Moss's New Media Medicine group at the Media Lab. There he develops tools that empower patients to educate each other through data and help accelerate biomedical research. Through a long-term engagement with a rare-disease community, Ian developed LAMsight, a patient advocacy tool for open-ended collection of self-report patient data. A spin-out project, the International LAM Registry, unifies clinical data from clinicians around the world to open up and consolidate patient records for more effective exploitation.

Ian is an advisor to a wide variety of startup companies, venture capital institutions and non-profit organizations; he holds over a dozen patents in semiconductor, telecommunications, and embedded software technology.

Peter Szolovits



Peter Szolovits is Professor of Computer Science and Engineering and head of the Clinical Decision-Making Group within CSAIL. His research centers on the application of AI methods to problems of medical decision making and design of information systems for health care institutions and patients. He has worked on problems of diagnosis, therapy planning, execution and monitoring for various medical conditions, computational aspects of genetic counseling, controlled sharing of health information, and privacy and confidentiality issues in medical record systems.

Peter Szolovits' interests in AI include knowledge representation, qualitative reasoning, and probabilistic inference. His interests in medical computing include Web-based heterogeneous medical record systems, life-long personal health information systems, and design of cryptographic schemes for health identifiers. He teaches classes in artificial intelligence, programming languages, medical computing, medical decision making, knowledge-based systems and probabilistic inference.

Prof. Szolovits is on the editorial board of several journals, has served as program chairman and on the program committees of national conferences, and has been a founder of and consultant for several companies that apply AI to problems of commercial interest.

Frank Moss



An entrepreneur and 25-year veteran of the software and computer industries, Frank Moss has spent his career bringing innovative business technologies to market. Most recently, he co-founded and is on the board of Infinity Pharmaceuticals, Inc., an early-stage cancer-drug discovery company doing innovative work at the intersection of technology and the life sciences. In addition, he chaired the advisory council for the creation of the Systems Biology Department at Harvard Medical School, where he remains an advisor.

During his career in the computer and software industries, Moss served as CEO and chairman of Tivoli Systems Inc., a pioneer in the distributed systems management field, which he took public in 1995 and subsequently merged with IBM in 1996. He co-founded several other companies, including Stellar Computer, Inc., a developer of graphic supercomputers; and Bowstreet, Inc., a pioneer in the emerging field of Web services.

He began his career at IBM's scientific center in Haifa, Israel, where he also taught at the Technion, Israel Institute of Technology. He later held various research and management positions at IBM's Yorktown Heights (NY) Research Center, working on advanced development projects in the areas of networking and distributed computing; and executive management positions at Apollo Computer, Inc., and Lotus Development Corporation.

He received a BS in aerospace and mechanical sciences from Princeton University, and both his MS and PhD in aeronautics and astronautics from MIT. His citations include Ernst & Young's Entrepreneur of the

Year award and Forbes Magazine's "Leaders for Tomorrow."

Henry Lieberman



Henry Lieberman has been a research scientist at the MIT Media Lab since 1987. His research interests focus on the intersection of artificial intelligence and the human interface. He directs the Lab's Software Agents group, which is concerned with making intelligent software that provides assistance to users through interactive interfaces. In 2001 he edited *Your Wish is My Command*, which takes a broad look at how to get away from "one size fits all" software, introducing the reader to a technology that one day will give ordinary users the power to create and modify their own programs. In addition, he is working on agents for browsing the Web and for digital photography, and has built an interactive graphic editor that learns from examples and annotation on images and video.

Lieberman worked with the late Muriel Cooper, who headed the Lab's Visual Language workshop, in developing systems that support intelligent visual design; their projects involved reversible debugging and visualization for programming environments, and developing graphic metaphors for information visualization and navigation. From 1972-87, he was a researcher at the MIT Artificial Intelligence Laboratory, and subsequently joined with Seymour Papert in the group that originally developed the educational language Logo, and wrote the first bitmap and color graphics systems for Logo. He also worked with Carl Hewitt on Actors, an early object-oriented, parallel language, and developed the notion of prototype object systems and the first real-time garbage collection algorithm. He holds a doctoral-equivalent degree (Habilitation) from the University of Paris VI and was a visiting professor there in 1989-90.