Proxies for Clinical Effectiveness in Genetic Information Visualization

Marijke Rijsberman
Interfacility
PO Box 620283
Woodside, CA 94062
+1-650-868-3432
marijke@interfacility.com

ABSTRACT
In this position paper, I describe a genome scan information visualization, whose effectiveness could be evaluated on the basis of intelligibility, user satisfaction, engagement, and behavioral and clinical outcomes. While clinical outcomes, being tied to customer health, present themselves as undoubtedly the best evaluation criterion, they are the most difficult to assess in the product development lifecycle in a startup environment. Practically measurable proxies have to be considered instead..

General Terms
Algorithms, Design, Reliability, Experimentation, Human Factors,

Keywords
Genome, Genetics, Test results, Information Visualization, Clinical outcomes, Evaluation criteria, Evaluation design, Information brokerage, Intelligibility, Customer satisfaction, Engagement, Behavioral outcomes.

1. INTRODUCTION
A startup company (hereafter referred to as GS startup), is currently in the throes of developing a "genome scanning" service for the general public, as a preventive medicine tool. The company's fundamental philosophy is to present consumers with information about their genetic predisposition for preventable diseases for which genetic associations have been established. (The definition of preventable diseases includes not only diseases that can be prevented altogether, but also diseases whose onset can be delayed or where early diagnosis can lead to more effective treatment.) The company hired me initially to perform user research and eventually to help design the risk assessment report that forms the core of its service.

GS startup essentially positions itself as the middleman between consumer and genetic laboratory. The customer buys the service online, and then receives a "spit kit" for the collection of a saliva sample, using a technology that determines the genetic sample, which is forwarded to the lab for analysis. The lab scans information at 1 million locations (called SNPs) on the customer's genome.[1] The results represent a (small) subset of the totality of genetic information the customer's genome contains. The lab records these results on a chip, which it sends to GS startup.

2. INFORMATION BROKERAGE
The real service GS startup provides is to filter and translate the vast quantity of information on the chip and make it useful to a lay person, that is, someone who is neither a genetic scientist nor a medical professional. The first step is to tie the information on the chip to the scientific literature that establishes associations between specific genotypic information with disease outcomes, typically expressed in odds ratios.

The preventable diseases filter simplifies the information to be conveyed to the customer to some extent—only a small subset of the million SNPs are relevant to this perspective. In other ways, the preventable disease filter makes the challenge more complex. By definition a preventable disease does not conform to the simple mental model that most lay people have of genetics, that of an on/off switch. According to this mental model, "on" means you have a particular gene and you will get the associated disease and "off" means you don't have the gene and will not get sick.

Preventable diseases are typically associated with multiples SNPs (in some cases more than 10 and the numbers are likely to grow). Each combination of alleles at each relevant SNP is associated with an odds ratio, which in itself is not a concept the average consumer is very comfortable with. What's worse, how the odds ratios for different SNPs combine has not been established conclusively. In other words, information brokering takes place in an environment of considerable scientific uncertainty

[1] The 1-million-SNP scan is currently the only commercially available genome scanning technology. A new, more complete methodology for genomic analysis has been released recently but is not yet available as a commercial service.

Among many other things, then, SG Company could tell its customer the following for any given disease on the list:

- what allele combinations the customer has for each of the associated SNPs
- what the odds ratios are for each of the SNPs
- what the prevalence, incidence, or lifetime risk is of the disease, providing a baseline for odds ratio calculations

3. PROTOTYPES AND MODELS

Early prototyping made it abundantly clear that a typical consumer in the target audience wants a simple result, much along the lines of standard medical testing results. Evidence of malignancy, infection, damage, or some such either is found or it is not. Alternatively, results fall within a certain acceptable range or they do not. The patient does not have to understand the test itself or how results are computed. While s/he may expect to be involved in the complex decision-making that ensues, the patient typically does not wish to understand such things as prevalence or calculate odds.

The same turned out to be true for a genome scan of this nature. In fact, representative customers evince no particular interest in their genes at all, only asking for a quick indication of where their health risks lie and then spending all their time trying to figure out what they can do to mitigate that risk. Of course, the bar goes up as the time the customer is willing to spend comes down.

In the face of this preference for simple results—a make or break consideration for product success—the GS startup science team developed a formula to estimate lifetime risk on the basis of the customer's results for all the relevant SNPs for for each of the diseases. That left two pieces of information to convey for each of the conditions in the panel:

- what GS startup calculates the likelihood to be that the customer will develop the disease
- whether the estimated likelihood is higher or lower than the average lifetime risk for the US population of the customer's gender

Three models for visualizing this information were bandied about in the wake of these developments. Each model uses a framework presenting lifetime risk, based on genetic information in 5 categories from low to high.[2]

3.1 Model A

Model A uses color to indicate what the customer's overall lifetime risk is, on the basis of an intuitive color association from green to orange to indicate which diseases to worry about the most. A rollover would show the actual risk number and the population average to compare it to. A simple filter makes it possible to see all conditions where the customer's estimated risk is above average or those conditions where it is below. See figure 1 for an illustration.

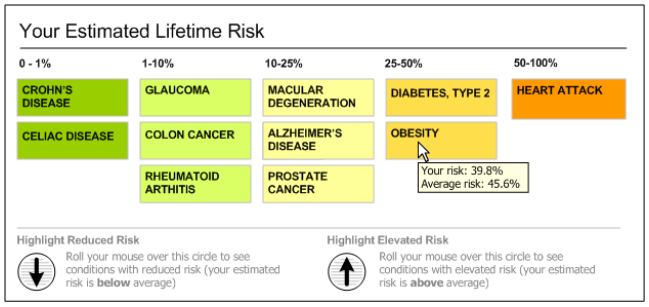

Figure 1: Model A – Color indicates risk

3.2 Model B

Model B maintains the risk "bins" but uses color to show whether the result for each condition is above, near, or below the population average. The risk numbers are now shown within the boxes, to make it possible for the astute customer to understand what the color means. A key becomes necessary to explain the significance of the color. Model B de-emphasizes overall risk in favor of feedback on "test results," and it introduces a new notion of "near" average that says some variation is not significant. The boundary between what is significant and what is not significant is drawn arbitrarily in this case, a significant drawback of the model.[3] See figure 2 for an illustration.

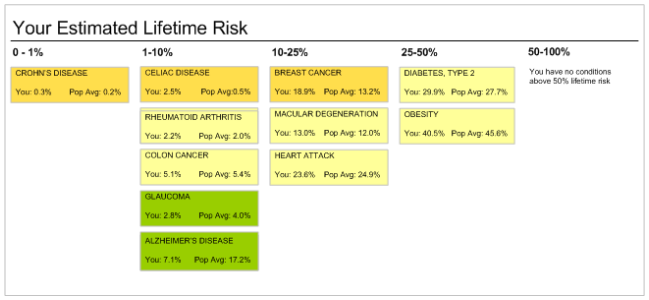

Figure 2: Model B – Color indicates elevated or reduced risk

3.3 Model C

Model C also uses the risk "bins," but uses color to indicate where the customer should focus his or her attention, despite the fact that the company could not come to agreement about what kinds of results warrant the most attention. Some people thought there is significance in absolutely high lifetime risk only (let's say, arbitrarily, anything above 25%), while other people thought there is also significance in relatively high risk (let's say, anything more than a fifth above average) even if the absolute risk is still low.

This approach would tell you to pay attention to your risk of getting celiac disease if it is three times higher than average, but still at only 1.5%, which is not the case for the approach that

[2] Personalized environmental information can be added into the calculation in future, to come up with a closer approximation of actual lifetime risk, but the company is not currently in a position to offer this feature.

[3] Other drawbacks include effects felt elsewhere in the report, where the results are represented as part of a navigation scheme, sorted from high to low. In model A the colors are neatly arranged, while in model B they appear in an odd jumble.

focuses only on absolute risk. [4] One of the drawbacks of Model C is that the significance of the color cannot be inferred from the display itself. In Model B, the numbers correlate to the colors, and the customer could infer the logic even though that eventuality is unlikely. See figure 3 for an illustration of Model C.

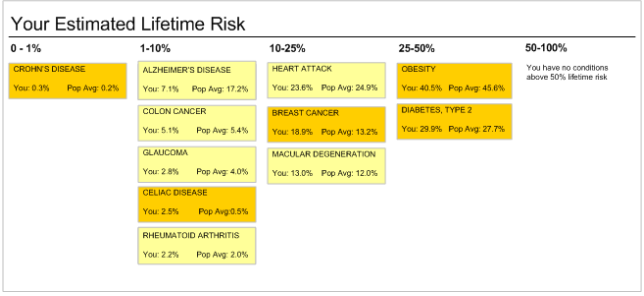

Figure 3: Model C – Color indicates suggested "focus"

4. THE EVALUATION CHALLENGE

Despite the lack of consensus, model C was chosen as the preferred beta implementation. The evaluation challenge then lies in designing an evaluation approach that can help determine whether model C performs as expected (and at least better than models A and B). The potential evaluation criteria include the following:

- Intelligibility (does the customer understand the results, a criterion in which errors play some role)
- User satisfaction (do customers feel that the experience lives up to their expectations)
- Engagement (do customers become more aware of health risks, do they talk to their doctors, do they undergo diagnostic testing, begin treatment regimens, etc.)
- Behavioral outcomes (do customers make changes in their behavior—most significantly, do the results drive a positive change in diet and exercise)
- Clinical outcomes (do customers have better health outcomes)

Given the mission of the company, clinical outcomes are clearly the best criterion for evaluating the effectiveness of the different models. However, of the options listed above, the clinical outcomes criterion is the most difficult and expensive to assess, requiring longitudinal studies that dip well below the practical horizon that bounds the purview of a startup company. Even engagement and behavioral outcomes (even if we only look at short-term behavioral changes, which is not ideal) would require AB testing in order to evaluate effectively, and AB testing is not feasible the beta timeline. Whether AB testing is feasible after beta is highly questionable. That means that we are thrown back on intelligibility and user satisfaction as the only criteria that can effectively be measured within the normal product development lifecycle.

While these criteria, being user-centered, strike me as a reasonable approach with respect to such products as consumer goods, here it would be important to know whether an approach that makes users happy is also the approach that makes them healthy. The dilemma illustrates rather painfully how the choice of evaluation criteria in a practical context is not bounded only by what we need to know but what we can realistically ascertain within the confines of standard business practice.

Conversely, the case illustrates how the choice of evaluation criterion has the potential to transform the product development lifecycle, in that it implies requirements for artifacts needed to measure whether the information visualization successful clears the bar.

[4] A third potential driver of "focus" would be to consider which risks are *easiest* to mitigate. GS company is not able at this time to provide an algorithm to capture an "easy-to-mitigate" driver of focus.