Apple Inc. announced HealthKit on June 2, 2014. The API was designed "to save, query, and interpret health data to storage of personal health information," allowing them to spend their time "keeping users happy and healthy". Apple’s efforts do not represent an isolated case. Numerous research prototypes and commercial tools provide the opportunity to collect, monitor, analyze, and make decisions based on health data that you collect about yourself. Pedometers such as Fitbit allow detailed tracking of physical activity. Electrodermal activity (EDA) sensors record levels of physical arousal. Promotional materials for fitness tracker Microsoft Band suggest that "this device can know me better than I know myself, and can help me be a better human". Similar developments can be seen in academic research. Numerous endeavors under the umbrella of personal informatics "help people collect personally relevant information for the purpose of self-reflection and gaining self-knowledge" while some systems focus on behavioral health indicators, such as physical activity, while others focus on chemical health markers, such as blood sugar.

Traditionally, such medical data were collected by, viewed in the context of, and interpreted alongside medical practitioners. However, devices embodying the emerging data-intensive approach to health and well-being represent a fundamental change to the location and usage of medical data. The advent of these systems raise numerous questions about the evolution of medical practice. Exactly what role(s) do we envision such technologies playing in a broader medical ecology? Where do health data tools sit in mediating, or perhaps disrupting, patient-clinician relationships? How do we design such tools to facilitate individuals interpreting their own health data outside of a clinical setting? Medical practitioners and health policy analysts offer various arguments. The field of HCI is uniquely positioned to augment such arguments with empirical evidence about these kinds of systems. While much prior work has focused on the use of personal informatics tools, relatively little research has considered how various design strategies might impact users’ crucial abilities to interpret detailed medical data.
actively and passively, a wide range of information about with a rich array of sensors to track and record, both using their pocket supercomputers laden with a rich array of sensors to track and record, both actively and passively, a wide range of information about

Individuals are using their pocket supercomputers laden with a rich array of sensors to track and record, both actively and passively, a wide range of information about themselves. For instance, seven in ten U.S. adults track a health indicator (blood pressure, mood, weight, blood sugar, sleep, etc.) for themselves or for a loved one [20]. These individuals believe there is great potential in using such collected data to positively impact their lived experience through greater self-knowledge. The research community has documented such individuals’ “lived informatics” practices [10,48] and how they collect and use such data to make changes in their lives [34]. Indeed, personal data-driven practices are transforming (and transformed by) the ways we understand and evaluate the individual in society at large [40].

Increasingly, individuals are tracking data that until very recently was only accessible in clinical and laboratory settings. In a very real way, lab tests are moving outside the laboratory, becoming not lab tests but simply additional data points self-tracked by everyday people. For example, the company Theranos performs inexpensive lab tests, and results are sent directly to both patient and clinician. Regulation in some regions is moving to make such practices more available. Indeed, in Arizona any person can now legally order any lab test themselves without first obtaining a doctor’s note or dealing with insurance eligibility [28]. The efforts of 23andMe provide an analogous example in genetic testing. Of direct relevance to the current work, in 2014 the US Food and Drug Administration (FDA) ordered 23andMe to discontinue its marketing, primarily citing serious concerns about test result accuracy and interpretation. “Two weeks after receiving the letter, 23andMe took steps to allay the FDA’s concerns. It continues to offer DNA testing and analysis, but no longer provides new consumers with genetic interpretations that relate to health” [25]. The FDA’s concerns resonate with Kay’s recommendations around improved weight scale design, perhaps the simplest and most pervasive of medical tests conducted by laypeople. “We encourage designers to think about how to design better scales given their inherent uncertainty and the properties of the data being collected” [31:409].

**Patient-Clinician Relationship**

The trend of lab tests moving outside the medical lab, and eventually onto laypeoples’ phones and becoming simply ‘tests,’ reflects and informs the changing style of relationships patients and the wider population have with their clinicians. Starting in the 1970s, a move from a simple ‘clinician as authority’ model has increasingly become one of ‘patient as consumer’ where there is an onus on the patient to be active managers of their own self-care [39]. With the rise of eHealth, we are in the era of the digitally engaged patient, “when lay people are advised that they should use digital technologies as part of patient engagement practices” [39:258–259]. One further iteration of this trend is the call to authorize the layperson to initiate and direct health behaviors, including laboratory tests - whatever the testing site. “Direct access to laboratory tests ushered in a new era of individual empowerment to take

**RELATED WORK**

**Health Apps and (Self-)Tracking**

In the United States, ownership of personal technology and its use is incredibly pervasive. Over 78% of people own desktop or laptop computers, 90% own cellphones, and 64% own smartphones specifically[2,15]. Globally, it is estimated that by 2020 80% of adults will have a smartphone [17]. Furthermore, 62% of Americans use their phone to look for health and medical information online [2].

The data-capture capabilities of these technologies are becoming increasingly extensive and sophisticated. Now-standard sensors on many mobile smartphones include GPS, accelerometer, compass, gyroscope, ambient light detection, proximity detection, dual microphones, and dual cameras [38]. Furthermore, adoption of sensor-laden wearable devices similarly continues to grow as effective and affordable commercial products (e.g. Fitbit Flex, Jawbone Up, MyBasis, Apple Watch, Android Wear, etc.) become available, e.g., for sleep tracking [9].

Individuals are using their pocket supercomputers laden with a rich array of sensors to track and record, both actively and passively, a wide range of information about

system called NutriPhone [33], which combines off-the-shelf blood testing strips, a custom hardware adapter, and computer vision algorithms to transform any smartphone into a micronutrient tester. Using vitamin B12 testing as a case study, this paper considers how best to design interfaces for presenting test results to users, as well as how users might interpret and act on those results. Specifically, we test different representational formats (numerical, graphical, verbal, etc.) both for users’ subjective preferences and for users’ ability to interpret the results accurately. Combining results from an online survey using participants from Mechanical Turk (N=249) and a laboratory study (N=27), we find that, in all but the baseline case, the majority of participants successfully interpret the test results correctly. Depending on the interface, accuracy ranged from 71.2% to 98.6%. However, users’ subjective preference for each interface did not closely align with their accuracy in interpreting the interface. Furthermore, many study participants expressed little confidence and doubted their interpretations, even when they were correct.

These findings offer mixed implications. On the one hand, they suggest that, by judiciously leveraging specific design elements, users can accurately interpret an interface showing the results of medical testing. On the other hand, users’ lack of confidence suggests such devices cannot and should not replace medical practitioners. We conclude by summarizing the positions both for and against giving medical data directly to patients, supplementing perspectives from prior work with data and results from our own experiments. Thus, this work contributes both recommendations for designing medical data feedback systems and evidence about the contexts in which it might be most (and least) appropriate to deploy such technologies.
control of our own health, to work with our physicians to prevent disease and detect it early, instead of treating illness after it has taken hold” [28]. In this way, clinicians become layperson-managed specialists, even simply a supporting cast in one’s own self-care. In this perspective, the above-mentioned Arizona legislation means that “Arizonans who want to know if they’re at [...] are empowered to engage with their physicians sooner” [28]. Furthermore, those who order these tests are envisioned to be from the general population. Just as the medical lab tests are moving out of the labs and becoming simply ‘tests’, the subject of these tests has moved from being a ‘patient’ in the doctor’s office and is now out in their daily life.

While for many a utopian ideal, the representation of the digitally engaged patient, empowered with mHealth apps and self-directed medical testing, may become problematic in several ways. First and most obviously, it assumes that end users have both the economic and technological capability to participate in digitally engaged patienthood [39]. Secondly, there are populations who are uncomfortable challenging their doctors as medical authorities, “particularly if they are elderly, from a socioeconomically disadvantaged or marginalised social group or very ill or suffering severe pain” [4,26,51] [cited in 39:263]. Third, while a patient may see digitally empowering health technologies both as a way to remain in contact with a doctor and (mediated by CMC) a support emboldening them to ask questions they could not in person [3], s/he may also see them as a way of avoiding doctor visits [3]. Finally, the responsibility of well-being self-management on the shoulders of the patient can be a significant burden. Not only can passive tracking feel intrusive and lead to a sense of being surveilled [1], but frequent manually-conducted tests and self-reports can be disruptive [52] and result in people using the systems far less frequently.

Layperson Sensemaking of Test Results
The trend of putting lab tests and their results in the hands of laypeople has raised some concern that patients may not accurately make sense of the data. Primary reasons include widespread low numeracy and health literacy in the US population at large [18,22]. Also, written information about the tests and their results are being provided at higher reading levels than most patients can manage [18] and in uninformative presentation formats [58].

An individual starting with lower numeracy is already at a disadvantage when trying to make sense of lab test results. Even if the numbers are intelligible, patients need more than numbers. “Clinicians mentally interpret test values in terms of categories of risk and thresholds for action that are unfamiliar to or even hidden from most patients. This contextual information provides a rich, colorful palette of meaning to clinicians, but most patients can see their results only in black and white terms—that is, as a good or a bad result” [45:2]. While visual displays of numerical health information have been found to aid patient understanding [37], individuals vary widely in their capabilities and preferences in understanding numbers, graphs, and other presentation formats in the context of health information interpretation [21].

Lower health literacy can mean a lack of understand not only about the markers being measured, but also regarding the meaningfulness and uncertainty associated with those measurements. It is not clear whether explicitly presenting such uncertainty to end-users will aid them in their task [as in 5] or not [as in 50], perhaps due to additional cognitive load or even by lowering confidence in the system. Focusing on in-home weight scales, Kay et al. [31] recommend avoiding false precision (e.g., tenths of a pound) in single-point, instantaneous feedback. Instead, they suggest that scale designers consider presenting weights in larger resolutions, such as 5 lb. increments, or focusing on relative trends rather than absolute values.

Finally, effective use of the data presented also greatly depends on the user’s motivation and questions when interpreting the data. Compared to basic questions such as “is my result abnormal,” test analyses could encourage engagement with more practical questions such as “does this mean I am at immediate risk?” and “has my value changed enough to make a difference?” [45:2].

Even in a one-on-one, face-to-face setting, clinicians can find it difficult to meaningfully convey the results of health tests and risk communication. In fact, in a commentary for the National Cancer Institute, Fagerlin et al. [18] distill many of the above aspects of effective risk communication, verbal literacy, graphical literacy, and numeracy, into ten straightforward recommendations for clinicians. These include using plain language, recognizing that risk communication is persuasive as well as informative, and considering conveying only the information most critical to patient decision-making, even at the expense of completeness.

STUDY CONTEXT
Given the technological developments surveyed above, sophisticated personal medical tests will be a reality for the general population in the near future. The important implications of these tests require informed design of the interfaces used to present test results for general use. Therefore, this study addresses the question “(how) should patients have direct access to medical data?” with an exploration of whether and how participants make sense of blood test results. Blood tests offer one example of personal medical tests that the general population will be able to conduct on their own, potentially without the guidance of a healthcare practitioner. As a case study, we present the NutriPhone system, a personal medical device which aims to address the problem of malnutrition [33].

More than two billion people worldwide suffer from micronutrient deficiencies, which places a significant burden on global development and economic progress. Unfortunately, while the problem is clear, most deficiencies
are chronic and largely asymptomatic until later in life. Thus, many who are at risk of serious complications remain unaware of their own vitamin status until it is too late [54]. This problem is compounded in resource-limited settings, such as developing regions [14] or low-SES populations [27], where the risks of malnutrition are highest and conventional bloodwork is unavailable due to a lack of medical infrastructure and trained personnel. NutriPhone is a smartphone- or tablet-integrated platform that can catch these deficiencies without a centralized hospital or a professional phlebotomist. This system will enable rapid personalized awareness of suboptimal nutrition through direct quantitative physiological feedback.

The system comprises three parts: a paper microfluidic test strip, a hardware accessory which clips around the smartphone or tablet camera, and a software app that guides the user through the test protocol, including acquiring and processing images, interpreting and displaying results, and storing data for long-term tracking (Figure 1).

The NutriPhone hardware accessory consists of a pair of battery-powered LEDs and focusing optics to allow for high-quality, focused images of the test strips while keeping the accessory thin and portable. The optical setup is held over the smartphone camera with an opaque 3D printed plastic enclosure, which blocks external light from hitting the sensor and ensures reproducible imaging independent of ambient lighting conditions. This allows the NutriPhone device to be used as a portable optical microscope that can capture magnified images of the test strip and analyze them via an image processing routine. The device accepts interchangeable paper test strips which can quantitatively measure a wide variety of micro- and macronutrients in whole blood with a single drop of blood from the fingertip.

The NutriPhone software performs a variety of functions throughout the testing procedure aimed at minimizing user errors, ensuring measurement accuracy, and displaying and storing the final results.

1. The NutriPhone app contains the standard operating protocol for each of the available test strips. Because a key focus of the device is minimizing the level of training required for use, the app guides the user through the appropriate procedure step-by-step with diagrams and timers. This reduces the likelihood of a mistake that could invalidate the results.

2. After a blood sample is added and the necessary time has elapsed, the app captures an image of the test strip using the smartphone camera. The test region of this strip is then analyzed and compared to a measured calibration curve, and the quantitative B12 concentration is calculated.

3. The measured B12 concentration is then stored in a local SQLite database, which can be used to track trends in the patient vitamin status over time and/or send results directly to the patient’s doctor for interpretation. The results are also displayed on the screen for immediate feedback to the user. The total test time, from finger prick to displayed results, is around 10 minutes.

METHODS
To explore designing to facilitate the interpretation of health data, we conducted two studies in parallel. In an online survey, participants evaluated a set of static prototypes. In a lab study, participants used an interactive prototype of NutriPhone and completed a debrief interview.

In both studies, we focused on users’ experiences making sense of health data related to a vitamin B12 test. Vitamin B12 is a nutrient that helps keep the body’s blood and nerve cells healthy, as well as being involved in making DNA. As B12 is naturally found in a wide range of animal products, most people in the US and other developed nations get all the B12 they need from the foods they eat. Those groups that may not get enough B12 include vegans, those with celiac or Crohn’s disease, those with pernicious anemia, and many older adults due to impaired intestinal absorption. B12 deficiency is associated with tiredness, weakness, loss of appetite, weight-loss, and anemia. Nerve problems, including nerve damage, can also occur. Both depression and osteoporosis have been linked to diminished vitamin B12 status. B12 is not known to cause any harm, and there is not a known “overdose” level, although B12 can interact or interfere with some medicines.

Despite its importance, doctors rarely test B12 in isolation or discuss B12 levels with their patients, since B12 deficiency is fairly uncommon in developed nations, such as the US [8]. Thus, study participants were unlikely to have prior knowledge of B12 and would thus need to rely more on the interface for interpreting and making sense of test results. Contrast this with more common health markers, such as cholesterol or BMI, where lay persons are

![Figure 1: The NutriPhone device, a system for point-of-care nutrition monitoring, consists of three parts: (1) a plastic hardware adapter that clips around the iPhone/iPad camera, (2) a disposable microfluidic test strip for blood sample analysis, and (3) a software app to acquire and process images, as well as interpret and display test results.](image-url)
more likely to have knowledge of what constitutes un/healthy levels.

**Online Survey – Comparing Interface Designs**

**Participants**

Amazon’s Mechanical Turk (MTurk) was used to recruit 249 participants (93 female), ranging from 18 to 66 years old. Most participants identified as American or European (presumably because the survey was in English), although 18 were from India and 3 from other Asian countries. All received $1.25 for their participation.

**Materials – Interface Designs**

We explored five designs for presenting test results.

**Number:** The B12 level was presented as a number. This design represents a baseline, presenting only the raw number that would result from a blood test.

**Color:** The B12 level was presented using a “stoplight” style color system, in which a healthy result was represented as green, and a low result was represented as red. These colors were drawn from a popular tool for measuring malnutrition [56]. The colors were tested for colorblindness and, except for individuals with deuteranopia (insensitivity to green light), were still distinguishable.

**Verbal:** The B12 level was presented using natural language to indicate whether the B12 level was healthy or low.

**Graph:** The B12 level was presented visually in the form of a graph, with a marker placed at the respective B12 level. The graph included anchors of “low” and “healthy” to provide orientation.

**Hybrid:** The B12 level was conveyed using a combination of all the color, verbal, and graph designs.

For each design, we created up both a healthy and a low/deficient mock-up. This arrangement gave us ten different feedback interface design variants (Figure 2).

**Procedure**

Participants completed an online questionnaire accessed through MTurk. After consenting to participate, they were then asked demographic questions such as age, gender, and citizenship. They were then shown a randomly selected B12 results presentations and were asked multiple-choice questions about that presentation (e.g., “Based on the result above, what is your B12 level?” with choices of “healthy,” “too high,” “too low,” “unsure,” or “not enough information to answer”). This process was repeated for two other randomly selected presentations. Participants then saw an attention check question asking to which vitamin the previous questions had pertained. Six participants failed the

<table>
<thead>
<tr>
<th>B12 Level</th>
<th>Number</th>
<th>Color</th>
<th>Verbal</th>
<th>Graph</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy (within reference range)</td>
<td>vitamin B12 results 501 ng/dl</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
</tr>
<tr>
<td>Low (below reference range)</td>
<td>vitamin B12 results 199 ng/dl</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
<td>vitamin B12 results</td>
</tr>
</tbody>
</table>

Figure 2: We tested five different feedback interface designs (Number, Color, Verbal, Graph, and Hybrid). For each, we included a variant with Healthy levels of B12 and a variant with Low levels of B12.
attention check and were removed from the analysis. At the end of the survey, participants were asked to rank the three designs they viewed in order of preference, from most to least favored. Finally, a free response option asked the participant what other information s/he would like included with the result.

**Lab Study – Interactive Prototype Debrief Interviews**

We conducted semi-structured interviews with participants after asking them to use an interactive prototype of the NutriPhone system. There were 27 participants in this study (20 female), who ranged from 18 to 45 years old. Thirty participants were undergraduate students who received course credit for their participation, and the remaining three were academic staff who volunteered their time. Interviews lasted approximately 10 minutes, were all conducted in-person, and were audio recorded and transcribed.

Participants were first asked to use an interactive prototype of NutriPhone on their personal smartphone, which they accessed via a link we provided at the start of the study. There were two versions of the interactive prototype; one version displayed a B12 level within the US National Institutes of Health (NIH) recommended reference range, and the other displayed a B12 level lower than that reference range [43]. Both versions represented the result as a marker on a color-gradient graph, based on [33], with the exact number located above the graph (Figure 3). The version that participants used was randomly assigned; 14 participants saw the healthy variant and 13 the low variant.

Under the B12 result were three clickable buttons that participants could view: “What is B12,” “How to Improve B12,” and “More Resources,” which displayed the B12 pages from the National Institute of Health, Wikipedia, and WebMD.

The app was introduced to participants as a way to “help people run blood tests on their own without a healthcare practitioner,” and they were asked to imagine that they had already completed the blood test, as described in the Study Context section above. They were then instructed to explore the app for as long as they wished. These use sessions lasted approximately five minutes.

Following their use of the system, we asked participants to describe their test result. Based on their result, we followed up with questions about how they understood (or did not understand) their test result, how they typically make sense of medical test results, and their overall impressions.

We analyzed these interviews using inductive qualitative methods drawn from grounded theory [24]. First, three of the authors reviewed each interview and independently took notes about key ideas. We then used affinity diagrams to organize these ideas into the themes we discuss below.

**RESULTS**

Overall, the results point in two directions. First, the majority of participants across both studies correctly made sense of the test results. However, the interfaces participants most preferred were not those with which they were most accurate in sensemaking. Furthermore, the majority of participants also expressed confusion, doubting their own interpretations. We first consider participants’ successful sensemaking, then address points of confusion.

**Sensemaking**

Our online survey found that the majority of participants correctly interpreted the test result. With the exception of the Number design, accuracy ranged from 71.2% (Color design with Low B12 levels) to 98.6% (Hybrid design with Low B12 levels). Table 1 presents accuracy results for the healthy B12 and low B12 variants of each design, as well as net accuracy for each design (the arrangement parallels the structure of Figure 2).

<table>
<thead>
<tr>
<th>B12 Level</th>
<th>Number</th>
<th>Color</th>
<th>Verbal</th>
<th>Graph</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>12.0%</td>
<td>78.9%</td>
<td>96.0%</td>
<td>75.3%</td>
<td>92.1%</td>
</tr>
<tr>
<td>Low</td>
<td>2.6%</td>
<td>71.2%</td>
<td>87.8%</td>
<td>81.5%</td>
<td>98.6%</td>
</tr>
<tr>
<td>Net</td>
<td>7.2%</td>
<td>75.2%</td>
<td>91.9%</td>
<td>78.5%</td>
<td>95.3%</td>
</tr>
</tbody>
</table>

Table 1: Survey participants’ accuracy for each feedback interface design. The top row shows accuracy with the healthy variant (i.e., within reference range), the middle row shows accuracy with the low variant (i.e., below reference range), and the bottom row shows net accuracy for each design. Showing B12 level as only a Number performs worst, while our Hybrid design, which combined color, verbal, and graph elements, performed best.

Only some of these differences in accuracy are statistically significant. First, the accuracies for each design space do not differ significantly between the Healthy and the Low variant, except for the Number design (Fisher’s exact test, p=0.030). Thus, we compare the overall designs rather comparing Healthy and Low variants separately. Compared to the baseline of presenting only a Number, all feedback interface designs we tested performed significantly better.
Lastly, a few participants came to the conclusion that the relationship between their numerical result and the generated somewhat more nuanced interpretations based on being “in” or “out” of the reference range (Participants explained, “I know that it was between 0 and 900, is that right? So 500 is in the middle...” Participant 103 noted that when compared to 900, their result of 199 “would not be a very high percentage.” Less commonly, participants made sense of the results on the lower and upper ends of the scale, respectively. For example, participant 211 responded “It was to the right of the scale, so I’m assuming good.” Participant 102 explained “It kind of seems like it’s low because it’s toward the end of the scale.” For at least two participants (P#211 and P#201), the graph was far more memorable to them than the numerical result was. This finding reinforces the importance of visual elements in sensemaking [21,22,37].

Participants also frequently drew on color to make sense of the result. The graph included a color gradient from red to green, and participants drew on common conventions of these colors to identify red as bad or unhealthy and green as good or healthy. Participants explained their reasoning with statements such as “It was in the green area so I think it’s good” (P#205) and “Well, it’s in the red zone, so I’d say that’s bad” (P#111). This result reinforces our choice to base the color gradient on a popular tool for malnutrition [56], and suggests its broader applicability in conveying health information.

Less commonly, participants made sense of the results interface by comparing the numerical result (either 199 or 501) to 200 or 900, which were benchmark numbers located on the lower and upper ends of the scale, respectively. For example, participant 103 noted that when compared to 900, their result of 199 “would not be a very high percentage.” Participant 206 explained, “I know that it was between 0 and 900, is that right? So 500 is in the middle [...] I was just like okay, it seems fine.” This approach mirrors how blood results are typically interpreted (i.e., comparing a numerical result to an established reference range). However, rather than combing to a binary determination of being “in” or “out” of the reference range, participants generated somewhat more nuanced interpretations based on the relationship between their numerical result and the anchor values.

Lastly, a few participants came to the conclusion that the result presented was low based on the inclusion of the “How to Improve B12” resource displayed below the result (P#102, P#112, P#207). As participant 207 explained, “For the additional resources it talked about a deficiency, so I think that maybe it’s bad.” No participant who was presented with a normal B12 result came to this conclusion, which may suggest that this type of evidence is not enough to overrule other (maybe more salient) types of information. Rather, the type of supplemental information provided can act as supporting evidence in sensemaking.

Collectively, these results help understand not only which design elements support better sensemaking but also why. The importance of the anchor points for our lab study participants explains in part why the Number design was difficult to interpret. Conventional interpretations for the Color and Graph designs likely aided participants in making sense of those variants. Finally, the Hybrid design supports sensemaking using different types of literacy (verbal, graphical, etc.) [21,22,37], thus increasing the likelihood that any given participant will correctly interpret it. However, accuracy with the Verbal design did not statistically significantly differ from that with the Hybrid design. Given the space constraints of displays on small devices, a verbal display may thus in some cases offer a viable option to a hybrid design.

Objective Accuracy vs. Subjective Preference
In addition to accuracy, our online survey participants also ranked the three variants they saw in order of subjective preference. Table 2 shows subjective rankings (left) compared to objective accuracy (right). These subjective rankings align closely with objective accuracy. Given the US Food and Drug Administration’s plans to regulate health-related smartphone apps [19], these results have implications for the kinds of evaluative criteria that should be included in the regulatory approval process.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Subjective Preference</th>
<th>Objective Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Hybrid</td>
<td>Hybrid</td>
</tr>
<tr>
<td>2nd</td>
<td>Graph</td>
<td>Verbal</td>
</tr>
<tr>
<td>3rd</td>
<td>Verbal</td>
<td>Graph</td>
</tr>
<tr>
<td>4th</td>
<td>Color</td>
<td>Color</td>
</tr>
<tr>
<td>5th</td>
<td>Number</td>
<td>Number</td>
</tr>
</tbody>
</table>

Table 2: Comparison of subjective preferences (left) vs. objective accuracy (right) with each of the ten feedback interface design variants. Rank correlation is high ($\rho = 0.90$).

Confusion
The above results suggest that users can effectively make sense of medical data. However, other aspects of our results suggest potential issues with direct access to medical data. Even though 25 of the 27 in-lab study participants were able to interpret the interface correctly, 20 participants expressed at least some level of confusion. When asked what their test result was, responses were often a variant of “I don’t know” (P#211) or “I have no idea what that means.”
One way in which participants hedged their responses was in terms of evaluating their result relative to the median. While there are recommended upper and lower limits B12 levels, “An increased vitamin B12 level is uncommon” [43]. Furthermore, high B12 levels do not cause toxicity and are a diagnostic criteria for few conditions [16,30]. However, many medical markers have health consequences for abnormalities both above and below the reference range, such as blood sugar levels [44]. As a result, some participants wondered whether an ideal result would be located in the center of the scale, rather than toward the highest end of the scale. As participant 110 remarked, “Usually in the middle is what I think is good for these kind of things. Obviously, for academic tests, the higher you get the better, but I feel like the median is for blood results or blood pressure or something like that...not too high or not too low.”

Participants also hedged their responses in terms of the potential influence of individual or demographic characteristics. Participant 203 wanted the system to tell her what “a healthy range would be, and maybe you could even put in your specific demographics, if it matters. Like what your gender and your height and your weight are.” Indeed, medical tests such as creatinine are interpreted with respect to personal characteristics, such as sex [42]. However, the NIH has only one set of recommendations for blood serum B12 level, and does not categorize based on demographic factors such as age, sex, and body weight for adults [43]. This point created confusion for ten participants, who expected these factors to influence the interpretation of their results.

In addition to expressing confusion about how to interpret the result, participants consistently wanted more interpretation and prescription about what the result meant. For instance, they asked whether the test result warranted further action, such as diet change, use of supplements, or a follow-up with a physician.

In terms of interpretation, some participants wanted to see an average or normal score/range to better situate their results. This desire was reflected in statements such as “I was hoping it would tell me what my goal was” (P#102) and “One thing I wanted to know was what was the average, what’s normal” (P#201).

Other participants wanted a very clear-cut, binary interpretation of the result that indicated whether their result was problematic. As participant 105 put it, “Should I get help, or should I not?”

Data from the free response question in the online survey yielded similar results. Participants wanted to know, for instance, “what healthy levels of vitamin B12 are,” “how to interpret the results,” and, most importantly, “whether it is a situation to be consulted with the physician.”

DISCUSSION

The results of our study provide mixed implications for whether or not (and if so, how) people should be given direct access to their medical data.

Promises of Direct Access to Medical Data

One very encouraging trend we saw was that participants were often able to correctly identify the B12 results that they viewed, as long as there was some context provided. In the online survey, accuracy rates for the presentations that included anything beyond just a number ranged from 75.2% to 95.3%. This is in spite of a lack of general knowledge about B12, as evidenced by the poor accuracy rates with the Number-only design (7.2%), and general concerns about low numeracy and graphical literacy rates which arise in discussions of interpreting medical data [21,37]. Similarly, 25 of the 27 in-lab study participants correctly interpreted their result, albeit with some prompting.

These findings support the argument that providing participants with direct access to their data can be useful, even when those data include less familiar medical markers. Providing opportunities for people to interact directly with their medical data may also enable them to make better informed decisions about their care [13], to have greater feelings of control over their treatment plan [45], and to receive their results in a more timely manner, which can be critically important for some treatment plans [45].

We saw glimpses of the power of giving people direct access to their medical care, as several participants described wanting to use NutriPhone to monitor their own health without the assistance of a physician. Some participants imagined using a tool like this to ease their nerves or act as a screening, while some others envisioned using it as an actionable guide for making lifestyle changes, such as modifying diet or taking supplements. Offloading this type of health maintenance to patients is attractive to physicians, patients, and insurance companies. Leveraging medical technologies has been noted as a way to manage the shortage of adult primary care physicians in the US [7].

Perils of Direct Access to Medical Data

While some evidence points to benefits of providing direct access to medical data, our results also highlight some disadvantages of providing this data directly.

Although the majority of our participants correctly interpreted the test result, many also found the results confusing. Twenty participants in the in-lab interactive prototype study expressed significant confusion or doubt when interpreting the results. Open-ended responses from the online survey suggest a similar desire for the guidance of a medical professional.
This confusion presents a clear problem for providing people with direct access to their medical data, especially as systems such as NutriPhone are able to analyze increasingly complex and/or meaningful medical markers. If people are unsure of their results, it undermines many of the benefits discussed above, as “data that are not understood will always remain data unused” [45:2]. As a result, physicians are mixed on whether direct access to medical results is a good idea, especially for abnormal results or for results of tests with significant consequences, such as cancer or HIV [53].

Many of our participants also wanted to see additional interpretation of the results, with some explicitly wanting the system to “diagnose” whether the result was healthy. Doing so goes beyond what is traditionally reported in a blood result (i.e., the patient’s value and a reference range) by attaching some meaning to that result. However, making a diagnosis is typically the domain of health care practitioners, and for good reason. Currently, the practice of medicine is frequently ambiguous and uncertain, even for physicians [23]. Test results must be carefully and contextually interpreted in ways not easily codifiable.

Furthermore, some of our participants stated that they did not have the requisite knowledge to interpret the results or would prefer to leave the interpretation to their physician. One study found that patients felt significantly less confused by lab test results if their physician had discussed what to expect prior to viewing the results [11]. Our results reinforce patients’ desire for such trained assistance in interpretation.

An Alternative: Medical Data as Mediator

Given these tensions, we offer a vision for the future of devices similar to NutriPhone to serve as tools to be used in conjunction with one’s physician, rather than substituting for the physician. That is, such systems would act as mediators between patients and clinicians.

For example, a person could complete a routine blood screening using NutriPhone in the hours before an appointment with their clinician. Immediately after the test, the results are available for the patient, but the results are also sent to the clinician. These results would be discussed during the appointment, allowing the clinician to interpret any abnormal findings and for the clinician and patient to agree on a treatment plan if applicable. If follow-up tests are appropriate, NutriPhone could be used for this type of monitoring.

This scenario preserves many of the advantages of direct access to medical data while mitigating the disadvantages. Patients would still be able to interact directly with their medical data, enabling them to be more informed about their health and empowering them to have a voice in their treatment options. If people conduct blood screenings prior to appointments it may allow for more focused visits and fewer follow-ups, as well as shorter turnaround times. However, the results would also be viewed by a healthcare practitioner who would then provide the interpretation of the results, recommend treatment or follow-up tests.

Some studies have shown that when patients believe that better knowledge of their bodies is achieved via self-monitoring devices they feel as if they are more in control, and this leads to greater security and reassurance. However, if the data they produce suggest that their health is suffering, or if these data conflict with their own subjective and phenomenological interpretation of their state of health and well-being, this can be unsettling and anxiety- or fear-provoking. For example, Hortensius et al. [29] found that some users of glucose monitors described technology as a ‘friend,’ bestowing peace of mind, confidence, freedom and certainty. Others, however, referred to it as a ‘foe.’ They disliked having to prick their finger constantly, and they felt ashamed, anxious, helpless, or frustrated by glucose readings that were not in the appropriate range. Another study investigated patients’ experiences self-monitoring their chronic obstructive pulmonary disease. The biometric readings these patients produced on their bodies, such as their oxygen saturation levels and lung function, were valued for their objectivity, their ability to uncover the mysteries of their own bodies. The patients responded emotionally to the numerical data they produced, feeling encouraged, more secure, or reassured when the numbers were in the acceptable range, but experiencing anxiety, depression, or fear of physical deterioration when their data exceeded this range [39].

LIMITATIONS AND FUTURE WORK

Several limitations should be considered when interpreting these findings. First, the results presented to participants were not their actual vitamin B12 levels, but rather pre-generated data. As a result, participants may not have reacted the same way that they would if interpreting their own results. On the other hand, participants have been demonstrated to display demand characteristics when participating in lab studies [46], wherein they try exceptionally hard to perform well at the given task. Thus, participants may have been even more confused if interpreting actual data in a more naturalistic setting. Future work should explore how people evaluate their own medical results under these circumstances.

Also, our scales displayed B12 as having no upper reference range limit. However, as discussed above, many medical tests have both upper and lower reference range limits, and future work should explore the sense-making process for these types of medical markers.

While the design elements tested here (colors, numbers, graphs, and words) offered significant differences, this study represents a partial exploration of this vast design space. Future work should consider other elements that might appeal to different kinds of literacies [21,22,37,58]. Furthermore, given the空间 constraints of small devices, such as smartwatches or the notification area of smartphones, alternatives to space-consuming hybrid
designs should be explored. Finally, design explorations should also be conducted for feedback of other types of medical results. Such work would be especially valuable if focused on data such as BMI where knowing simply whether or not one’s value in “within normal limits” may not be sufficient.

Lastly, this study captured participants’ sensemaking at one point in time. Future work would benefit from considering potential learning or habituation effects arising from viewing subsequent tests.

CONCLUSION
Medical technology is changing rapidly. Numerous devices and systems place increasingly more health and medical information directly in the hands of patients. These technological developments present the possibility of a digital imperative [39,55,57], wherein such systems become integrated as part of medical care. In many countries, medical records are already largely electronic. Technologies similar to NutriPhone may, despite potential issues of interpretability, become a similarly integral part of medical care. When that day comes, HCI researchers and practitioners will need to be prepared to design tools that not only are comprehensible but also effectively mediate between doctor and patient.

Results such as those presented here will be invaluable to designing such systems. For example, we find that hybrid feedback designs, which include multiple different representational modalities (colors, words, etc.), result in the most accurate understandings of medical data. This finding suggests interfaces that tap into different literacies will likely be most effective at presenting medical data.

Furthermore, results suggest that such technologies should likely be designed not for independent use by patients but as an additional means of mediating the doctor-patient relationship. Prior work has made this argument on the basis of medical practice or policy recommendations [18,45]. This paper provides a valuable contribution to that debate in the form of empirical data about the design and interpretation of feedback interfaces for medical data.

REFERENCES
2. Monica Anderson. 2015. 6 facts about Americans and their smartphones. Washington, D.C.
8. Centers for Disease Control and Prevention. 2013. Guidelines for the Evaluation of the Nutritional Status and Growth in Refugee Children During the Domestic Medical Screening Examination. Atlanta, GA.


45. Maurice O’Kane, Danielle Freedman, and Brian J. Zikmund-Fisher. 2015. Can patients use test results effectively if they have direct access? The BMJ 350, h673. http://doi.org/10.1136/bmj.h673


