Evaluating Expert Curation in a Baby Milestone Tracking App

Ayelet Ben-Sasson
University of Haifa
Haifa, Israel
asasson@univ.haifa.ac.il

Eli Ben-Sasson
Technion
Haifa, Israel
eli@cs.technion.ac.il

Kayla Jacobs
Technion
Haifa, Israel
kayla@cs.technion.ac.il

Elisheva Rotman Argaman
Technion
Haifa, Israel
eargaman@campus.technion.ac.il

Eden Saig
Technion
Haifa, Israel
edens@cs.technion.ac.il

Figure 1: Expert curation of milestone texts in babyTRACKS, showing syntactically novel milestone texts on the left and a semantically novel milestone text on the right.

ABSTRACT

Early childhood developmental screening is critical for timely detection and intervention. babyTRACKS[1] is a free, live, interactive developmental tracking mobile app with over 3,000 children's diaries. Parents write or select short milestone texts, like *"began taking first steps"*, to record their babies' developmental achievements, and receive crowd-based percentiles to evaluate development and catch potential delays.

Currently, an expert-based *Curated Crowd Intelligence (CCI)* process manually groups incoming novel parent-authored milestone texts according to their similarity to existing milestones in the database (for example, *"starting to walk"*), or determining that the milestone represents a new developmental concept not seen before in another child's diary. CCI cannot scale well, however, and babyTRACKS is mature enough, with a rich enough database of existing milestone texts, to now consider machine learning tools to replace or assist the human curators. Three new studies explore (1) the *usefulness* of automation, by analyzing the human cost of CCI and how the work is currently broken down; (2) the *validity* of automation, by testing the inter-rater reliability of curators; and (3) the *value* of automation, by appraising the "real world" clinical value of milestones when assessing child development.

We conclude that automation can indeed be appropriate and helpful for a large percentage, though not all, of CCI work. We further establish realistic upper bounds for algorithm performance; confirm that the babyTRACKS milestones dataset is valid for training and testing purposes; and verify that it represents clinically meaningful developmental information.

[1]Formerly Baby CROINC, CROwd INtelligence Curation [1].

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300783

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**;

KEYWORDS

Curated crowd intelligence, crowd wisdom, early childhood development

ACM Reference format:
Ayelet Ben-Sasson, Eli Ben-Sasson, Kayla Jacobs, Elisheva Rotman Argaman, and Eden Saig. 2019. Evaluating Expert Curation in a Baby Milestone Tracking App. In *Proceedings of CHI Conference on Human Factors in Computing Systems Proceedings, Glasgow, Scotland Uk, May 4–9, 2019 (CHI 2019),* 12 pages.
https://doi.org/10.1145/3290605.3300783

1 INTRODUCTION

Tracking early childhood milestones (like *"first smile"* and *"crawling"*) improves pediatric well-being and developmental outcomes. Healthcare guidelines recommend regular professional screening of children under 3 years of age to prompt intervention and maximize its benefits early in development [9]. Nonetheless, in practice, the large majority (70%) of children with developmental conditions go undiagnosed at early wellness checkups [18], due to poor access to high-quality healthcare [8], inconsistent use of standardized screening tools by medical providers [34], and a significant time gap between a parent's first concern and first consultation with an expert [13]. Parents of young children are increasingly relying on the internet and social media for information about child development, yet they are flooded by information, much of which is irrelevant to their own children [3, 12, 25, 31, 38].

To lower barriers to developmental screening and empower parents to be more involved with tracking their children's progress in between appointments with healthcare professionals, we developed babyTRACKS (formerly Baby CROINC, CROwd INtelligence Curation [1]), a free website and mobile app[2] which helps parents assess their children's development through crowd-based statistics curated by experts. While traditional child development evaluation tools are closed, static checklists of milestones with specific ages at which they should be achieved (e.g., Ages and Stages Questionnaire (ASQ) [4]), babyTRACKS offers a model for early developmental screening in a digital media world, and positions parents as reliable reporters about their children. It is an open system which enables parents to use their own words to describe their observations; recognizes new, relevant developmental milestones emerging from societal changes (e.g., *"started swiping at smartphone"*); and allows non-linear trajectories that require flexibility in age ranges (e.g., walking before crawling).

Up to this point, the bulk of the work to curate and classify the novel milestones has been manually conducted by trained staff members with backgrounds in child development. We have now reached a point in our platform's maturity, and in the richness of our user-authored milestones database, where we naturally consider whether and how to automate some or all of that work. A critical step before automation is to check whether and how it is possible to do so well.

This pre-automation step is highly relevant beyond this specific platform and beyond early childhood development. This is particularly the case for other crowd-based projects which, like ours, require a very high level of accuracy in classification with significant consequences from errors (in our case, potential for unnecessary worry on the one hand, or false reassurance and subsequent treatment delay on the other).

Paper Organization

First, we review the literature, and then describe the babyTRACKS system with especial focus on the expert curation process. Next, we present a series of three new studies, each addressing different questions about automation preparation:

(1) Analyze the cost of the milestone curation process currently being performed by human experts—*would automation be useful?*
(2) Assess the reliability and reproducibility of the curation task, by examining the performance of non-experts on the classification task—*would automation be valid?*
(3) Examine the "real world" relevance of the curated milestone concepts in babyTRACKS, as assessed by child development experts—*would automation be valuable?*

Finally, the Discussion section brings together the insights from these studies and sets the stage for the opportunities and constraints ahead in designing machine learning methods for our platform and others like it.

2 RELATED WORK

Health/well-being tracking apps number in the thousands, with many focusing specifically on child development. Various applications allow parents of young children to track their child's feeding and sleeping schedules as well as their physical growth (e.g., Text4Baby [15, 24], Trixie Tracker [29]). The MyPreemie app [11, 17] and Estrellita app [21] specifically support parents of premature babies. In another platform, CUE [14], parents were trained to track seven motor milestones over time using a hand-held computer. @BabySteps [26, 36] shares babyTRACKS's goal of early developmental screening by sending parents a set of text messages or Twitter messages describing a developmental milestone relevant to the child's age (based on traditional evaluations), soliciting a parental response as to whether the milestone has been reached. babyTRACKS is unique for

[2]The babyTRACKS app is available on the web at http://www.babytracks.org and can can be downloaded on Google Play at https://play.google.com/store/apps/details?id=com.babycroic.croinc and on Apple's App store at https://itunes.apple.com/us/app/smart-baby-diary/id1185899162 .

its support of personalized text entries for developmental tracking and its quantitative feedback, in line with the goals of parent education and the democratization of knowledge.

Our work further builds upon several HCI areas: the digitally-engaged patient, personal lived informatics, crowdsourcing, and expert curation.

The digitally-engaged patient describes the active involvement of patients (or, in our case, their parents/caregivers) in self-monitoring themselves as part of preventative health [28, 37]. Various information and communication technologies facilitate health self-management [6]. PatientsLikeMe is the most active and studied online health self-management platform for individuals with chronic medical conditions. Patients' engagement in digitally self-monitoring symptoms has been associated with better health outcomes, efficient communication with medical providers [5, 42], reduction in healthcare costs [10], and the accumulation of aggregated data valuable for research [32].

Personal lived informatics [33] and the "quantified self" [37] have become integral parts of many people's lives. Self-trackers such as FitBit and Zeo enable real-time continuous health measurements such as heart rate and sleep quality. At a personal level, many individuals are continuously collecting and interpreting data about themselves; at a population level, unprecedented amounts of data are aggregating, offering new possibilities for health science [28, 37]. The "quantified baby" [39] concept has been applied to baby monitoring technologies. Evidence from a wearable monitoring device points to both positive and negative impact of baby quantification upon motherhood knowledge, self-competence and practice [39]. Tracking young children's developmental milestones is part of this new parenting role. Unlike passive sensor data collection, however, tracking development requires active observation and recording.

Crowdsourcing offers new opportunities for digital health research by generating data that is larger, broader in representation, faster to obtain, and more accessible [32, 41]. Thus far, crowdsourcing has been applied to many health domains including diagnostics, epidemiology, genomics, mental health, and nutrition (see surveys [32, 37, 41]. Crowdsourcing health can be situated within a broader revolution known as "New Power" [22]—open, generated and motivated by the crowd, and channeled through its growth and distribution. Our project meets all three main trends observed in crowdsourcing health research: non-professionals conducting science activities, recruiting participants online, and active health management using the internet [37]. While we focus here on estimating the quality and cost of curating early childhood development crowd data, our approach is relevant and easily applicable to crowdsourced clinical data in other areas too, like geriatrics and mental health.

Expert curation forms the basis for building clinical knowledge repositories. Domain experts classify health-related texts and develop knowledge databases which can guide clinical decision-making (e.g., [19, 27, 40]). Research mining genetic associations with psychiatric disorders demonstrated that careful expert curation workflow and a dedicated dashboard led to high inter-curator reliability [19]. Furthermore, expert curation is of importance for training machine-learning algorithms to support clinical predictions (e.g., [2, 30]). For instance, in a recent study [2] experts classified early autism markers mentioned in parental internet queries in order to facilitate automated initial autism screening based on internet queries.

Expert curators are often needed to balance the quantity and quality of crowd-generated data. Human involvement in training machine learning algorithms is of importance in the case of health informatics data which are often not standardized, noisy, involve rare events and missing data [23]. Interactive digital health platforms which rely on crowd data, as babyTRACKS does, often require experts to monitor the inputs to facilitate meaningful outputs. Chan and colleagues [7] demonstrated how expert versus inexperienced curators can significantly improve crowd ideas. Nonetheless, expert curation has two costs: an economic cost (human time, money, training); and a possible accuracy cost (errors, biases). Expert curation is not scalable and can become a bottleneck. Quantifying its pros and cons is necessary to define the opportunities for integrating automated machine learning tools.

3 BABYTRACKS AND THE CROWD CURATED INTELLIGENCE (CCI) PROCESS

babyTRACKS is a free early childhood development tracker available on the web and as iPhone or Android mobile apps[2], empowering parents' understanding of their young children's development and encouraging improved interaction with healthcare providers. A dynamic interaction takes place between user activity (e.g., adding milestone texts to their children's diaries), crowd data (e.g., milestones and ages entered by others), and expert curation (e.g., aggregation of milestone texts, defining canonical milestone concepts).

As shown in Figure 2, in babyTRACKS, the parent creates a developmental diary for his or her child by recording age-dated milestone texts, like *"smiles at people"* at five weeks old or *"stacking blocks"* at 14 months old (Figure 2). For each recorded milestone, the system provides the parent with statistical information associated with that milestone text, including the child's personal percentile achievement (percentages of children in the system who achieved the same milestone(s) at earlier or later ages). Parents can view these personalized percentile statistics for any milestone that has been reported in at least ten other children's diaries.

In addition, a developmental report summarizes the child's milestone percentiles within any particular developmental area (like social, gross motor, speech, etc.), including a graph of the child's milestone percentiles over time.

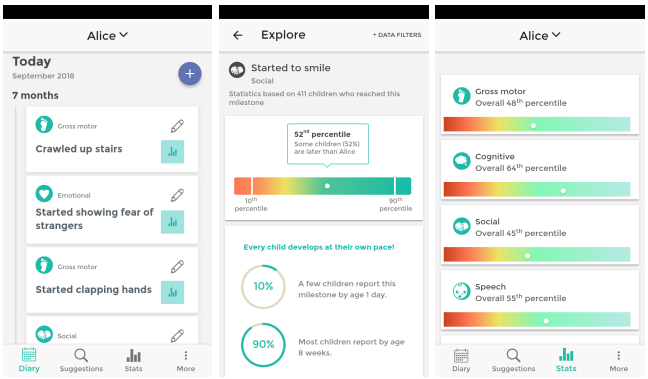

Figure 2: Screenshots from a child's developmental diary in the babyTRACKS mobile app. Left: Example diary. Middle: Exploring statistics for the *"started to smile"* milestone, showing that the child's reported age of 2 months was evaluated to be average (neither delayed nor early) at the 52^{nd} percentile, computed with respect to a sample size of $n = 411$ babyTRACKS children. Right: summary report of overall milestone percentiles in various developmental areas.

Previous babyTRACKS research [1] provided preliminary indication for the validity of the system's crowd-based percentiles. These percentiles differentiated between boys and girls, as well as preterm and full-term babies, in a manner similar to that observed by traditional, valiated developmental tests. Furthermore, babyTRACKS percentiles significantly correlated with age norms provided by the U.S. Centers for Disease Control and Prevention (CDC) [16] for corresponding milestone concepts.

Adding Milestones to Diaries

Parents have several ways of adding milestones to their children's diaries, as illustrated in Figure 3:

(1) Entering their own **original milestone text**;
(2) Starting to type text and then selecting an existing milestone from an **autocomplete** list; and
(3) Selecting from a list of **suggested milestones**, recommended by the system based on the child's personal characteristics, like age.

The last two methods automatically group the child's new milestone with milestones previously reported by other parents, but the first currently requires human expert involvement, the focus of this paper.

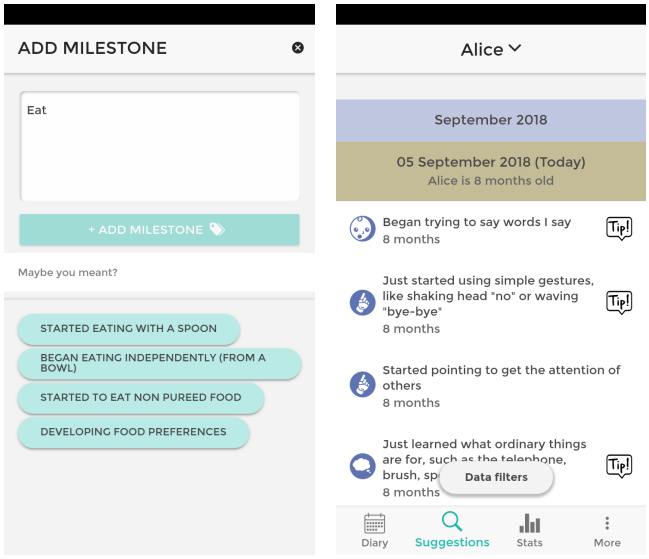

Figure 3: babyTRACKS interface for adding a new milestone text to a child's diary, with three possible methods. Left: Users can enter their own original milestone text (1), or select from an autocomplete list below (2). Right: Suggested milestones recommended by the system (3).

The Crowd-Curated Intelligence (CCI) Process

Crowd Curated Intelligence (CCI) is a behind-the-scenes expert curation process that enables babyTRACKS to provide crowd-based percentile statistics while preserving the option for parents to use their own words to describe their children's milestones and contribute to the kinds of milestones tracked by the system. Curators *merge* conceptually-similar milestone texts into unified canonical milestone concepts so that information from multiple children can be pooled together to provide statistical comparisons. Previous work about babyTRACKS [1] demonstrated that without CCI, only 12.80% of milestone concepts would contain enough data to provide users with significant statistics.

Syntactic novelties are user-authored milestone texts that are new to the system, but with semantic meaning similar to an existing canonical milestone concept. For example, if a user entered *"Sophia is toddling already!"*, a curator groups this milestone with an existing canonical milestone concept associated with other semantically-similar but syntactically-different milestone texts, like *"began taking first steps"* and *"he started walking"* (Figure 1).

If, however, a semantically-similar canonical milestone concept does not exist for a new milestone text, it is a *semantic novelty* representing a new, unique developmental concept which were not encountered by the system before, such as *"walks holding on to table"*. The curator creates a new canonical milestone concept for it, optionally lightly edits

the text to improve the phrasing (e.g., *"walks holding on to table"* to *"walking holding on to furniture"*), and categorizes it as part of one or more developmental areas (e.g., gross motor, fine motor, cognitive, social, language). In prior research [1], we described the increase over time of semantic novelties, starting from an initialization of 252 canonical milestone concepts[3] and growing slowly to 618 in the present system.

CCI work involves taking user-generated content and putting it into a meaningful context [1]. It is conducted by a staff member with a background in child development who has been trained for the task by the first author, a researcher specializing in child development. The supervisor also conducts periodic reviews of milestone text classifications to ensure quality and provide additional feedback.

Curation in babyTRACKS often requires domain-specific knowledge. Determining semantic similarity at times goes beyond simple linguistic matching of incoming ideas, and requires comparison of the reported age of the milestone texts to that in the system, analysis of the developmental areas involved, and understanding of the behaviors described in a canonical milestone concept. For example, a new milestone text *"went down the stairs independently"* at age eight months may raise questions like, does this refer to crawling or walking down stairs? Based on the young, pre-walking age, the curator can merge the milestone with the canonical milestone concept *"crawled down stairs independently"*.

Contemplating CCI Automation

While expert curation is very valuable, it does come with a cost: Maintaining high-quality expert curation requires recruiting and training curators, and ensuring they are constantly coordinated with each other.

Previous work [1] showed that each user-created milestone takes 5 minutes of curation time on average, and the cost-per-milestone remained relatively constant as the system scaled up. This means that a 100x larger system will cost 100x more curator hours, which is not feasible as our project continues to grow. Additionally, due to the expert review necessary, parents experience a short time lag (typically between several hours to a few days) between entering novel milestone texts and seeing their children's developmental percentiles for those milestones, which diminishes their user experience.

At the system's start a few years ago, CCI automation was not possible. We had no existing database of different milestone texts corresponding to the same canonical milestone concept, and only half of the current set of canonical milestone concepts.

[3]The initial list was drawn from the early childhood developmental milestones published by the U.S. Centers for Disease Control and Prevention (CDC), based on two developmental guideline books for caregivers by the American Academy of Pediatrics [20, 35].

At this point in babyTRACKS's maturity, naturally the question arises: can CCI be automated? Completely or partially (as time-saving suggestions for curators)? Is our database mature enough? Reliable enough? How much would automation help? Before hastily implementing a machine learning algorithm, we first step back to assess the system's readiness for automation, and what constraints and potential may be involved.

4 STUDY SERIES

Having reviewed the babyTRACKS system and the CCI process, we will now describe our three studies, exploring whether automation can be (1) useful; (2) valid; and (3) valuable.

babyTRACKS Database

Our studies are based on babyTRACKS database as of late October 2017, built up over several years of user and curator activity since 2015:

Children. babyTRACKS had 3,208 children, first registered by their parents at an average age of 9.28 months (median = 4.74 months, SD = 1.05 years), with 90% under the age of 2 years.

Milestones. Counted together, users added a total of 26,880 milestones to their children's diaries. These were grouped into 618 canonical milestone concepts (for example, *"started walking"* vs. *"counts to ten"*), and included a total of 2,749 unique user-authored milestone texts (for example, the distinct milestone texts *"started walking"* and *"began to walk"* are two different milestone texts for the same canonical milestone concept). Note there was a "long tail" of concepts; the most-popular 300 canonical categories cover over 96% of diary milestones.

The mean number of diary milestones per child, after accounting for some deletions, is 8.41 (median = 5.00, SD = 12.21). The mean age of the child when the milestone was achieved was 6.00 months (median = 3.42 months; SD = 7.72 months). Milestone texts were short, with a mean length of 7.34 words (median = 6 words; SD = 3.56 words; range 1–34 words).

CCI curation log. Expert curators review and process new milestone texts as they are added to users' children's diaries, through a behind-the-scenes curator-facing software interface. All of the CCI operations performed through this interface—6, 916 to date—are automatically logged, and by analyzing them we can understand how curator time is spent inside babyTRACKS. Some aspects of the experts' work are not covered (for example: training, or consulting colleagues over the phone), but since the majority of curation time is spent inside this interface, log analysis is a good proxy for understanding the overall CCI activity.

Study 1: Automation Usefulness: The Human Cost of Curation

The goal of the first study was to understand the different factors contributing to the human cost of performing CCI, and to assess the cost-reduction potential of automation. To do this, we analyze the flow through the manual CCI process.

Study 1: Methods. The analysis in Study 1 relies on babyTRACKS's CCI curation log. For each user-generated milestone text, the distinction between syntactic and semantic novelties was made by monitoring changes to its associated canonical milestone concepts. The milestone text was considered a syntactic novelty if it was assigned to an existing canonical milestone concept, and a semantic novelty otherwise (thus initializing its own new canonical milestone concept, to which future milestone texts could potentially be later assigned).

Study 1: Results. 92% of milestone additions occur through interface channels not requiring further human involvement (like autocomplete or milestone suggestion lists), and 8% of milestone text additions are original texts authored by users, requiring CCI processing. Of these, 83% are semantically similar to milestones that already exist in the system (7.0% of total milestone text additions), and the remaining 17% represent semantic novelties which were previously unknown to the system (1.5% of total milestone additions). See Figure 4 for a complete breakdown.

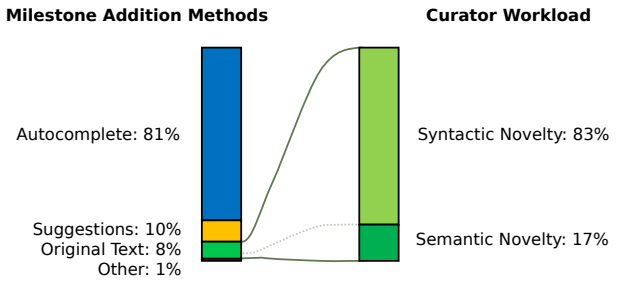

Figure 4: Flow through the CCI process. Users mostly add milestones to their children's diaries using the autocomplete interface (81%) or suggestions (10%). Original texts (8%), authored by users and previously unseen by the system, are processed by expert curators as part of the CCI process. Of these, 83% turn out to be semantically similar to canonical milestone concepts that already exist in the system, and the remaining 17% are semantic novelties.

While accounting for a relatively small percentage of milestone additions, original texts led (as hoped) to extensive syntactic diversity amongst milestone concepts. As shown in Figure 5, canonical milestone concepts had quite a few unique texts, with both a mean and median of over seven and a maximum of 48.

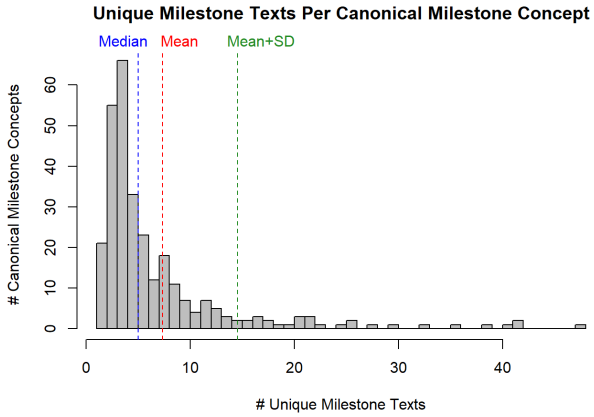

Figure 5: Syntactic diversity amongst canonical milestone concepts. Each concept (appearing in at least ten children's diaries, the system's minimum requirement for showing percentile statistics) had, on average, over seven different unique user-authored milestone texts associated with it.

Study 1: Discussion. We found that 98.5% of the milestone texts added by users were semantically similar to other canonical milestone concepts that already existed in the system, highlighting the low prevalence of semantic diversity among babyTRACKS's users, alongside high syntactic diversity in the number of unique milestone texts per canonical milestone concept.

In addition, even though automatic milestone suggestion mechanisms in the user interface have been shown to be helpful in reducing curator load (by prompting users to select existing milestone texts), expert curators still encountered existing canonical milestone concepts in more than 80% of the new milestone texts. Given that babyTRACKS does not yet employ much language processing mechanisms, this suggests that advancing towards an automatic solution for the semantic similarity problem may lead to significant gains, in both cost reduction and accuracy.

Study 2: Automation Validity: Inter-Rater Reliability of Curation

In our second study, we focused on the reliability and reproducibility of our curation process amongst human curators.

At the heart of CCI is a classification task: as detailed above, short texts written by parents about their young child's developmental milestones are either *merged* with existing canonical milestone concepts or defined as *new* concepts. Our goal was to determine the reliability and validity of this manual

classification to date: is it a subjective, inexact process, or rather an objective, reliable process with consistent results across human curators?

Only the latter outcome can enable consideration of potential future automation efforts on our dataset, as if even human reviewers frequently disagreed strongly with each other, an algorithm could not hope to achieve any real level of validity when trained and evaluated against this dataset. Additionally, understanding the level of disagreement between humans sets an upper bound for the eventual algorithm's evaluation metric goals.

This study simulated the manual curation process amongst several non-experts with basic child development knowledge, comparing their individual curation decisions with each other and with the babyTRACKS staff's previous curation decisions.

Study 2: Methods. We recruited participants meeting the following criteria:

(1) The mother of a child aged 1–6 years;
(2) Without a professional/educational background related to early childhood development (e.g., psychology, occupational therapy, education);
(3) No experience with babyTRACKS or other child development trackers; and
(4) Native English speaker.

These requirements were intended to ensure basic but not specialized familiarity with child development, as common to lay mothers who are not child- or healthcare experts.

Data was collected from ten mothers[4], though one was omitted due to extreme outlier performance[5]. The remaining nine mothers had a range of 1–4 children (M=2.33, SD=1) aged between 6 months to 9.67 years, of which at least one child was 27 months old or younger. Mothers were 28–33 years old (M=31.22, SD=1.56), held some sort of post-secondary education (from college diplomas through Ph.D. studies), and all but one were currently employed. All mothers were born in English-speaking countries (USA or Canada) and had immigrated to Israel 7–17 years prior to the study (M=11.67, SD=2.95).

Study 2: Design. The study participants were each given the same list of 100 milestone texts, selected randomly from the thousands authored by babyTRACKS users, which had been already been previously manually categorized by staff into canonical milestone concepts. To understand the overall system, participants also received a reference table (see excerpt in Figure 6) of the babyTRACKS canonical milestone concepts[6]. For each one of the 100 milestone texts, participants were instructed to either MERGE it into an existing canonical milestone concept appearing in the reference table (a syntactic novelty), or to declare it NEW (a semantically novel concept).

Canonical Milestone Concept	Milestone texts	Developmental area(s)	Median child age	Child age range
turns head toward sounds	-follows my voice on to both sides by turning his head -track after noisy moving objects -established turning head toward sounds	cognitive language	31 days	0 - 144 days (0 - 5 months)
started pointing to get the attention of others	-is pointing to get the attention of his brother -started to point to get daddy to look at her -pointed at cat to get grandma's attention -just learned pointing to show others something interesting	nonverbal communication	280 days	112 - 382 days (4 - 13 months)

Figure 6: Excerpt of the canonical milestone concept reference table.

Participants were given tips for finding similar canonical milestone concepts in the reference table, including:

- Searching within the reference table for the milestone text's main keywords and synonyms thereof.
- Considering that the milestone might contain language errors in spelling or grammar.
- Comparing the milestone's reported child age and the canonical milestone concepts' median child ages and/or typical child age ranges.

Participants were also encouraged to not necessarily pick the very first reasonable canonical milestone concept match, but rather to continue searching to see if there was a better match. If they did not find any matching canonical milestone concept, they were instructed to only then mark the milestone as NEW.

Study 2: Results. First we compared the participants' answers to babyTRACKS's answer. Participants' answers ranged from 60% to 73% agreement (mean = 67.11%, SD = 4.34%), meaning that their MERGE or NEW answer matched with

[4]Two participants originally recruited were excluded as they did not complete the task, and two others were recruited in their place. The participants were reimbursed for their time with a gift card.

[5]While each participant had different answers and rates of correctness and agreement, participant 1's was a clear outlier, as easily visualized in Figure 7. When the participants' answers were compared to babyTRACKS's answer, the answers of the other nine participants ranged from 60% to 73% agreement, while Participant 1's answers was a mere 42%, more than six standard deviations below the mean. When comparing the participants' answers with each other (and including babyTRACKS's answer as one of the participants), the difference became even more pronounced: The other participants' agreement rates with the highest-agreement answer ranged from 73% to 81%, while Participant 1's agreement rate was 50%. In light of these differences, Participant 1's answers were omitted from all other analyses.

[6]This reference table purposefully excluded the canonical milestones concepts for the milestones designed in this study to be marked as NEW by the participants, as well as the "long tail" of more esoteric concepts appearing only rarely in diaries—from which the 100 milestone texts were purposefully *not* drawn—resulting in a list of 287 concepts.

the professional curators' answer. When examining the errors that the participants made, we classified them in three categories:

(1) NO NEW: babyTRACKS's answer was NEW but the participant chose to merge the milestone text with another existing canonical milestone concept;
(2) NO MERGE: babyTRACKS's answer was to merge the milestone text with a canonical milestone concept but the participant chose to instead write NEW; and
(3) WRONG MERGE: babyTRACKS and the participant both chose to merge the milestone text, but the participant merged the milestone text with a different canonical milestone concept than the one chosen by babyTRACKS experts.

The most common error type was WRONG MERGE (mean = 15.56%, SD = 3.04%), followed by NO NEW (mean = 12.67%, SD = 3.24%), followed by NO MERGE (mean = 4.67%, SD = 3.53%). When examining agreement levels per milestone text, the average agreement rate with babyTRACKS's answer was 67.11%, with a standard deviation of 33.22%.

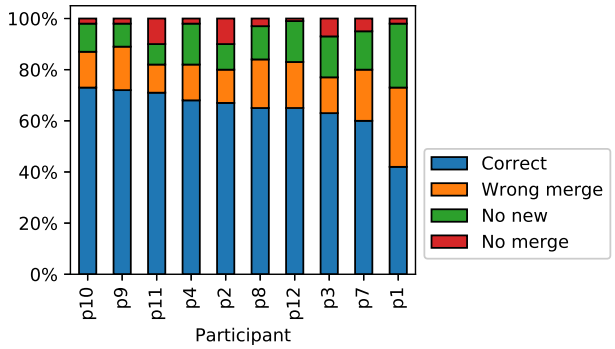

Figure 7: Participants' agreement with babyTRACKS (in percentages)

We then analyzed the data with babyTRACKS curators' answer included as another pseudo-participant, and examined the highest-agreement answers between all participants. Participants' agreement rates with the highest-agreement answer ranged from 73% to 81% (mean = 76.7%, SD = 2.50%). The percentage of milestone texts for which the highest-agreement answer was also babyTRACKS's answer was 76%. When examining agreement levels per milestone text, the average highest-agreement rate per milestone text was 76.87%, with a standard deviation of 21.98%.

Error Analysis. We studied 24 milestone texts for which the answer that received the highest agreement was not babyTRACKS's expert curators' answer. When calculating the agreement percentage, babyTRACKS's answer was included as a fellow pseudo-participant.

Highest agreement on these milestone texts ranged from 22% to 89%. The milestone texts fell into two groups: those with a clearly incorrect highest-agreement answer (meaning that babyTRACKS's answer was clearly preferable to the highest-agreement answer), and those for which the highest-agreement answer, though different from that given by babyTRACKS, was arguably correct. All in all, 14 highest-agreement milestone text answers were categorized as arguably correct, and ten highest-agreement milestone text answers were categorized as clearly incorrect. Of note, from an agreement level of 78% and higher, all the milestone texts were marked as having an arguably correct answer.

When conducting the error analysis, five distinct types of errors emerged among the 14 "arguably correct" milestone texts. These included cases in which the text could be categorized under either canonical milestone concept accurately (six texts), cases in which the text was vague (two texts), cases in which the "mistake" is a more accurate choice than babyTRACKS's answer (two texts), cases in which the highest-agreement canonical milestone concepts and babyTRACKS's canonical milestone concept are synonymous (two texts), and two cases of an error in the system—one in which babyTRACKS's canonical milestone concept was poorly defined and the other in which babyTRACKS's canonical milestone concept was missing from the reference table.

The ten "clearly incorrect" texts can be grouped into three distinct types. These included cases in which the participants probably stopped searching too early (four texts), cases which reflected the participants' lack of developmental knowledge (four texts), and cases in which the cause of the error was unclear (two texts).

Study 2: Discussion. When curators disagreed with each other or with babyTRACKS, it is clear that relatively few conflicts stemmed from a general misunderstanding of the system or the curation process; rather, most reflected specific points of debate regarding particular texts, needing a more nuanced assessment from an expert reviewer.

However, in most cases, babyTRACKS's categorization held true even among lay mothers; the participants' categorization of most (76%) of the texts matched the categorization in babyTRACKS's system. The level of agreement between participants was also encouragingly high, at 76.7% agreement.

We had planned to repeat the experiment with another cohort of participants with expert-level domain knowledge had our results been less robust, but this seems unnecessary: judging by these results, the curation was well-agreed upon between different judges to a high level, despite their lack of expert child development knowledge or training. A minority

of milestone texts did need the input of an expert with more advanced training.

We thus feel confident that our system's database, with its existing categorization of milestone texts into the various canonical milestone concepts, provides a valid, objective, high-quality gold standard dataset which can be used to train and test machine learning algorithms for future automation efforts. Importantly, when evaluating the performance of such algorithms, we know to aim for agreeing with human curators around 76% of the time (not 100%), since that is the level of agreement that humans can reach with each other.

Study 3: Automation Value: Real-World Relevance of Milestones for Child Development Evaluation

Having assessed, in the previous study, whether babyTRACKS canonical milestone concepts are reliable and reproducible categories, we now evaluate the clinical significance, usability, and "real world" value of these milestones for child development.

Study 3: Methods. Three experts with clinical backgrounds in child development were asked to independently rate canonical milestone concepts for two qualities:

(1) *Importance:* does having a record of this milestone's achievement aid in evaluating a child's developmental progress, or is it a superfluous piece of information?
(2) *Clarity:* is the real-world meaning of the milestone text understandable, or is the category too vague?

Both qualities were rated on a separate four-point scale (from "very important" to "very unimportant" and from "clear" to "very unclear," respectively).

The experts received the text of the milestone concept and the median age of the children with that milestone concept in babyTRACKS. We restricted our study to the 300 most popular canonical milestone concepts (each appearing in at least ten babyTRACKS children's diaries, and together covering over 96% of babyTRACKS diary milestones) to remove the "long tail" of more esoteric milestone concepts.

Study 3: Results. Overall, the three experts rated most milestone concepts as "very important" or "important" for 62%, 88%, 93% (respectively) of the concepts. "Clear" or "somewhat clear" ratings were assigned to 86%, 89%, 95% (respectively) of the milestone concepts.

Figure 8 breaks down expert rater agreement and disagreement. On the importance scale, 13% of concepts were rated identically by all three experts, 73% rated identically by two out of the three experts, and only 13% were rated differently by each expert. On the clarity scale, 36% of concepts were rated identically by all three experts, 52% rated identically by two out of the three experts, and only 13% were rated differently by each expert. For both importance and clarity ratings, fewer than 28% of concepts had rating disagreement that was moderate (a gap of two on the four-point scale) or extreme; the great majority had no or minor (a gap of one on the four-point scale) disagreement.

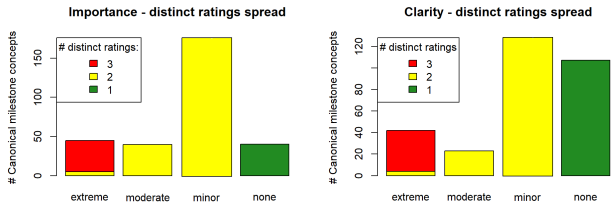

Figure 8: Spread of expert ratings of milestones' importance and clarity respectively. For both, only 28% had disagreements that were moderate or extreme.

Study 3: Discussion. We found high expert ratings of both importance and clarity for the great majority of milestones agreement between the experts, and high agreement between experts. These results give us confidence that most of the milestones in babyTRACKS reflect relevant, real-world information useful for child development tracking and assessments, as opposed to frivolous or irrelevant information.

We hypothesize that the un-evaluated "long tail" of rare canonical milestone concepts might contain a higher proportion of unimportant ones, as reflected by their lack of popularity amongst users.

5 DISCUSSION

babyTRACKS is a personal lived informatics platform enabling parent-driven developmental tracking of young children, harnessing the transfer of "New Power" [22] to parents with crowd-based percentiles. This paper discusses the interaction between crowd-based data and expert curation in providing insights for crowdsourcing health research to increase the quality of quantified-self data [37] in other fields.

Potential for Curation Automation

Each of our three new studies addresses a critical aspect of whether automating behind-the-scenes Curated Crowd Intelligence (CCI) work can be done with usefulness (Study 1), validity (Study 2), and meaning (Study 3):

Study 1 investigates how much of present system's curation is manual, and of that, how much is related to syntactic novelties (more easily addressable by an automated system) vs. semantic novelties (likely needing to remain manual). Without Study 1, it's unclear how much human work automation could potentially save, and thus whether it would be worth implementing.

Study 2 asks if there is reasonable human consensus on curation work ripe for automation attempts. Without Study

2, it's unclear whether curation is objective and reproducible, and whether we have a gold-standard dataset available for training and evaluating a valid automated algorithm. (For example, when babyTRACKS first started, there was no such dataset of user-authored milestone texts, and no avoiding manual human curation of incoming texts.)

Finally, Study 3 addresses whether even well-done curation—whether human or machine—has real-world, clinically meaningful insights for the relevant field of child development. Without Study 3, it's unclear what our work's real-world significance is, regardless of the curation quality (again, whether human or machine).

The results of the three studies establish that automation would save considerable curator effort (up to 83%, corresponding to the percentage of milestone texts which are syntactic novelties, semantically similar to existing milestones), that there is indeed human consensus on curation, and that the data is clinically meaningful. Together, they provide a powerful argument for automation's possible power and justify future work in that direction.

Limits on Curation Automation

Automatically processing the remaining 17% of semantically novel milestones will require more sophisticated approaches. Processing these milestones often requires more time, knowledge about child development, thought for determining their developmental distinctiveness, to compose a representative canonical milestone concept for the milestone and to link it with relevant developmental areas. These may require a continued integration of human curation, albeit in a reduced capacity than currently required by the system.

Additionally, Study 2's establishment of human curation agreement of around 76% gives us an upper bound for machine learning performance goals; due to the nuanced clinical nature of the subject, we cannot expect an algorithm to perform better than humans on this task, who achieve much less than 100% agreement.

Even if automatic curation cannot completely replace human involvement, however, it may be able to significantly speed up the process through offering, for each novel milestone text, suggestions of "top" canonical milestone concept suggestions for the curator to choose from, instead of obligating them to search through the entire list of 600+ (and growing) concepts.

Non-Expert Curator Possibilities

Study 2 demonstrated that with minimal training, non-experts (mothers of young children with practical experience but without formal backgrounds in child development) can broadly agree upon the curation process decisions. This shows the potential for training a computer program to deal with incoming milestones based on the pre-existing fine-grained curation, and also indicates the possibility for training or harnessing non-experts (perhaps even babyTRACKS users themselves) to conduct any remaining manual curation, integrating another potential crowdsourcing component into our system [37].

Nonetheless, experts will likely not be fully replaceable in curating semantic novelties and in supervising and reviewing the system periodically. The highly sensitive nature of child development data will likely require the integration of interactive machine learning algorithms which rely upon ongoing expert review and training [23]. The potential risks of "quantified baby" platforms include false reassurance or increased parental worry [38], necessitating a low tolerance for errors.

Conclusions

babyTRACKS is a free early childhood development tracker which allows parents to use their own words to describe their children's growth and contribute to babyTRACKS's repository of milestone concepts, while still providing parents with crowd-based statistical insights relevant to their children. To date, there has been a need for extensive behind-the-scenes manual expert curation, to build up a database of milestone texts and to ensure low error rates of associating conceptually-similar texts with each other. Our work demonstrated we now have a high-quality, meaningful, valid dataset on which to build machine learning algorithms to help automate this curation process to allow the system to continue to grow more scalably.

ACKNOWLEDGEMENTS

We thank the babyTRACKS team members for their contributions to the system design and maintenance: Gal Agmon, Moriah Anochi, Tal Bussel, Shir Har-Noy, Rotem Malinovitch, Daniel Moran, and Naama Tzur. Research was funded by the European Research Council (under grant agreement 240258), the Israeli Science Foundation (grants 1501/14 and 1435/18), the US-Israel Binational Science Foundation (grant 2014-359) and the Hiroshi Fujiware Cyber Security Research Center at Technion.

REFERENCES

[1] Ayelet Ben-Sasson, Eli Ben-Sasson, Kayla Jacobs, and Eden Saig. 2017. Baby CROINC: an online, crowd-based, expert-curated system for monitoring child development. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, New York, NY, USA, 110–119.

[2] Ayelet Ben-Sasson and Elad Yom-Tov. 2016. Online concerns of parents suspecting autism spectrum disorder in their child: content analysis of signs and automated prediction of risk. *Journal of Medical Internet Research* 18, 11 (2016), e300.

[3] Jay M Bernhardt and Elizabeth M Felter. 2004. Online pediatric information seeking among mothers of young children: results from

a qualitative study using focus groups. *Journal of Medical Internet Research* 6, 1 (2004), e7.

[4] Diane D Bricker, Jane Squires, and Linda Mounts. 1999. *Ages & stages questionnaires: aA parent-completed, child-monitoring system.* Paul H. Brookes, Baltimore, Md.

[5] Jed R Brubaker, Caitlin Lustig, and Gillian R Hayes. 2010. PatientsLikeMe: empowerment and representation in a patient-centered social network. In *CSCW'10; Workshop on research in healthcare: past, present, and future.* ACM, New York, NY, USA, 1–5.

[6] Jorge Calvillo, Isabel Román, and Laura M Roa. 2015. How technology is empowering patients? A literature review. *Health Expectations* 18, 5 (2015), 643–652.

[7] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, San Francisco, CA, USA, 1223–1235.

[8] Pamela Y Collins, Beverly Pringle, Charlee Alexander, Gary L Darmstadt, Jody Heymann, Gillian Huebner, Vesna Kutlesic, Cheryl Polk, Lorraine Sherr, Andy Shih, Dragana Sretenov, and Mariana Zindel. 2017. Global services and support for children with developmental delays and disabilities: Bridging research and policy gaps. *PLoS Medicine* 14, 9 (2017), e1002393.

[9] Bright Futures Steering Committee, Medical Home Initiatives for Children With Special Needs Project Advisory Committee, et al. 2006. Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics* 118, 1 (2006), 405–420.

[10] Roberto De Vogli. 2011. Neoliberal globalisation and health in a time of economic crisis. *Social Theory & Health* 9, 4 (2011), 311–325.

[11] Mia Wechsler Doron, Emma Trenti-Paroli, and Dana Wechsler Linden. 2013. Supporting parents in the NICU: A new app from the US, 'MyPreemie': A tool to provide parents of premature babies with support, empowerment, education and participation in their infant's care. *Journal of Neonatal Nursing* 19, 6 (2013), 303–307.

[12] Jodi Dworkin, Jessica Connell, and Jennifer Doty. 2013. A literature review of parents' online behavior. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 7, 2 (2013), article 2.

[13] Katherine D Ellingson, Margaret J Briggs-Gowan, Alice S Carter, and Sarah M Horwitz. 2004. Parent identification of early emerging child behavior problems: predictors of sharing parental concern with health providers. *Archives of Pediatrics & Adolescent Medicine* 158, 8 (2004), 766–772.

[14] Kate Ellis-Davies, Elena Sakkalou, Nia C Fowler, Elma E Hilbrink, and Merideth Gattis. 2012. CUE: The continuous unified electronic diary method. *Behavior Research Methods* 44, 4 (2012), 1063–1078.

[15] W Douglas Evans, Lorien C Abroms, Ronald Poropatich, Peter E Nielsen, and Jasmine L Wallace. 2012. Mobile health evaluation methods: the Text4baby case study. *Journal of Health Communication* 17, sup1 (2012), 22–29.

[16] Centers for Disease Control and Prevention. 2013. Learn the signs. Act early. Program. www.cdc.gov/ActEarly

[17] Gramham's Foundation. 2017. MyPreemie. http://grahamsfoundation.org/mypreemie-app/

[18] Frances P Glascoe. 2000. Early detection of developmental and behavioral problems. *Pediatrics in Review* 21, 8 (2000), 272–280.

[19] Alba Gutiérrez-Sacristán, Àlex Bravo, Marta Portero-Tresserra, Olga Valverde, Antonio Armario, MC Blanco-Gandía, Adriana Farré, Lierni Fernández-Ibarrondo, Francina Fonseca, Jesús Giraldo, et al. 2017. Text mining and expert curation to develop a database on psychiatric diseases and their genes. *Database* 2017 (2017), bax043.

[20] Joseph F Hagan, Judith S Shaw, and Paula M Duncan. 2007. *Bright futures: Guidelines for health supervision of infants, children, and adolescents.* Am Acad Pediatrics, USA.

[21] Gillian R Hayes, Karen G Cheng, Sen H Hirano, Karen P Tang, Marni S Nagel, and Dianne E Baker. 2014. Estrellita: a mobile capture and access tool for the support of preterm infants and their caregivers. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 3 (2014), 19.

[22] Jeremy Heimans and Henry Timms. 2014. Understanding 'new power'. *Harvard Business Review* 92, 12 (2014), 48–56.

[23] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[24] Voxiva Inc. 2017. Text4Baby™. https://www.text4baby.org/

[25] Kaylyn Khoo, Penny Bolt, Franz E Babl, Susan Jury, and Ran D Goldman. 2008. Health information seeking by parents in the Internet age. *Journal of Paediatrics and Child Health* 44, 7-8 (2008), 419–423.

[26] Julie A Kientz, Rosa I Arriaga, and Gregory D Abowd. 2009. Baby steps: evaluation of a system to support record-keeping for parents of young children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA, 1713–1722.

[27] Jessica JY Lee, Wyeth W Wasserman, Georg F Hoffmann, Clara DM van Karnebeek, and Nenad Blau. 2018. Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. *Genetics in Medicine* 20, 1 (2018), 151.

[28] Deborah Lupton. 2013. The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Social Theory & Health* 11, 3 (2013), 256–270.

[29] Ben MacNeill. 2017. Trixie Tracker™. https://www.trixietracker.com/

[30] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the future-big data, machine learning, and clinical medicine. *The New England Journal of Medicine* 375, 13 (2016), 1216.

[31] Lars Plantin and Kristian Daneback. 2009. Parenthood, information and support on the internet. A literature review of research on parents and professionals online. *BMC Family Practice* 10, 1 (2009), 34.

[32] Benjamin L Ranard, Yoonhee P Ha, Zachary F Meisel, David A Asch, Shawndra S Hill, Lance B Becker, Anne K Seymour, and Raina M Merchant. 2014. Crowdsourcing-harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine* 29, 1 (2014), 187–203.

[33] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal tracking as lived informatics. *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1 (2014), 1163–1172.

[34] Nina Sand, Michael Silverstein, Frances P Glascoe, Vidya B Gupta, Thomas P Tonniges, and Karen G O'Connor. 2005. Pediatricians' reported practices regarding developmental screening: do guidelines work? Do they help? *Pediatrics* 116, 1 (2005), 174–179.

[35] Steven P Shelov and Robert E Hannemann. 1993. *Caring for Your Baby and Young Child: Birth to Age 5. The Complete and Authoritative Guide.* Education Resources Information Centre, US.

[36] Hyewon Suh, John R Porter, Alexis Hiniker, and Julie A Kientz. 2014. @ BabySteps: design and evaluation of a system for using twitter for tracking children's developmental milestones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, NY, USA, 2279–2288.

[37] Melanie Swan. 2012. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research* 14, 2 (2012), e46.

[38] Anne M Walsh, Kyra Hamilton, Katherine M White, and Melissa K Hyde. 2015. Use of online health information to manage children's health care: a prospective study investigating parental decisions. *BMC Health Services Research* 15, 1 (2015), 131.

[39] Junqing Wang, Aisling Ann O'Kane, Nikki Newhouse, Geraint Rhys Sethu-Jones, and Kaya de Barbaro. 2017. Quantified Baby: Parenting and the Use of a Baby Wearable in the Wild. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 108.

[40] Zhuoran Wang, Anoop D Shah, A Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway. 2012. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 7, 1 (2012), e30412.

[41] Kerri Wazny. 2018. Applications of crowdsourcing in health: an overview. *Journal of Global Health* 8, 1 (2018), 1–20.

[42] Paul Wicks, Michael Massagli, Jeana Frost, Catherine Brownstein, Sally Okun, Timothy Vaughan, Richard Bradley, and James Heywood. 2010. Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research* 12, 2 (2010), e19.