Challenges in Developing Temporal Health Status Representations with Self-Tracked Data

Adrienne Pichon
ab3886@cumc.columbia.edu
Columbia University, Department of Biomedical Informatics
New York, NY, USA

Lena Mamykina
lena.mamykina@columbia.edu
Columbia University, Department of Biomedical Informatics
New York, NY, USA

Noémie Elhadad
noemie.elhadad@columbia.edu
Columbia University, Department of Biomedical Informatics
New York, NY, USA

Abstract

Endometriosis is a complex and poorly understood chronic condition with highly variable symptoms. While personal informatics tools support self-tracking, few leverage artificial intelligence (AI) to generate actionable insights for care. This research applies a digital phenotyping approach to characterize weekly health statuses using self-tracking data from the Phendo app — a research platform co-designed with endometriosis patients. Applying an unsupervised probabilistic mixed-membership model to week-level data, we generate temporal health status phenotypes that capture severity-based illness dynamics. We validate the learned phenotypes then evaluate them with a user study. We document various challenges with their alignment with participants' health status assignments. Findings highlight the importance of human-centered AI that enhances user autonomy and aligns with lived experiences. Future work will be needed to refine the phenotypes by incorporating additional details, such as medication usage and personal narratives from journal entries, ultimately improving AI-enabled tools for chronic disease management.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Smartphones**; • **Applied computing** → **Health informatics**.

Keywords

health status, chronic illness, self-tracking, personal informatics

ACM Reference Format:
Adrienne Pichon, Lena Mamykina, and Noémie Elhadad. 2025. Challenges in Developing Temporal Health Status Representations with Self-Tracked Data. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 5 pages.

1 Introduction and Background

Chronic conditions burden those who must live with and manage them [13, 15, 17, 20, 33], leading to a significant amount of work for patients and their care teams [4, 28]. Further, gaps in medical knowledge persist, complicating diagnosis, management, and treatment [21]. To mitigate these burdens, researchers have developed various personal informatics tools — technologies designed to collect, integrate, and analyze personal data to support reflection and decision-making — to support care for different diseases, support treatment, help reach patient goals, and improve quality of life [6, 9, 25]. With the proliferation of such personal informatics tools across health contexts and purposes, individuals managing chronic conditions often generate a considerable volume of personal health data about their illness experience that is available to support care [10]. With the rise of artificial intelligence (AI), there are even more opportunities to support care of chronic disease [29].

There is a particular opportunity to apply AI in the context of poorly understood, complex, enigmatic conditions, especially where the experience of illness varies greatly from individual to individual. While there are some intelligent systems to capture illness data and reflect on health-related data in these contexts (e.g., tools for identifying triggers in enigmatic diseases, such as irritable bowel syndrome (IBS) [14, 27], rosacea [3], and migraine [26]), few go beyond logging and reflecting on data or provide computational features to use the data to support care.

This research focuses on endometriosis, an inflammatory chronic, multi-factorial, and systemic condition estimated to affect 6-10% of women[1] of reproductive age [34]. In this burdensome chronic condition, the types and severity of symptoms vary greatly from individual to individual and over time. Endometriosis remains enigmatic [2], with substantial gaps in knowledge about the disease, leading to a lack of established medical guidelines [1]. Because endometriosis is poorly understood, has no biomarker for diagnosis, has no cure, and treatment is complex, individualized, and often ineffective, monitoring and care for the condition remains challenging [5, 11].

In complex and poorly understood diseases, it can be difficult to characterize individuals' health status, understand what is going on with their health, and communicate about the experience of illness with a care team [16, 22, 32]. Our prior work has detailed these challenges in the context of endometriosis [24]. In this research, we create and evaluate computable representations of illness states to help individuals characterize their health status and that may be used in intelligent systems. We ask the following Research Question: *Can a digital phenotyping approach aggregate individual-level data*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).

[1]Here, we reference endometriosis as a condition that impacts "women." While imperfect, the use of this term is important because "women's health" is under-studied and often stigmatized precisely *because* they are women's concerns, and stripping this label could obscure this problem (as also argued by Grimme et al. [12]). At the same time, we recognize that intersex people, non-binary individuals, and transgender men, for example, may also have endometriosis. We acknowledge that not all people who are impacted by "women's health" issues are women.

to enable interpretable representations of health status in the context of a complex enigmatic condition?

2 Methods and Materials

2.1 The Phendo App: A personal informatics tool for endometriosis

The Phendo app [7, 8] is a mobile research app that was developed in partnership with endometriosis patients [18, 19] to capture the real-world experience of the disease. The app enables users to catalog the day-to-day signs and symptoms, self-management activities, and other lived experiences of endometriosis outside of the clinic.

2.2 Phenotype development

Model. To address the particular challenges of these heterogeneous data and the complex, uncertain illness context, we rely on unsupervised probabilistic methods, similar to the approach taken by Urteaga and colleagues [31], who used data from the Phendo app for phenotyping individuals with endometriosis. This mixed-membership model is a specific type of generalized low-rank model, which is well-adapted to generating interpretable phenotypes. However, rather than aggregating an entire user's record for individual-level phenotyping, here, we extend this work to create temporal phenotypes, by aggregating each user's data by week. An interpretable, temporal representation (i.e., being able to represent an individual's timeline of health experiences as a dynamic mixture of phenotypes over different weeks) is likely to be suitable to real-world interventions. In the context of this research, the phenotypic profile characterizes the health status (comprised of characteristics across domains of illness) at a particular time for a particular person.

Data. We use data from the Phendo app. Eligibility requirements include: self-reported diagnosis of endometriosis and at least one self-tracking entry (with a minimum of five data points). All available self-tracking data are aggregated by user-week. Each user-week is described by a vector of vectors, where dimensions are represented as counts by specific item. Each domain represents a related set of tracking questions, mapped to meaningful responses. The final dataset includes data from a total of n = 11,852 users, who have tracked an average of 4.3 weeks. The dataset includes data from 51,187 user-weeks, with an average of 52.6 moments each.

Development. Development of the phenotypes is iterative. To determine the best hyperparameters to use for the phenotyping model, we experiment with held-out data that is split 80/20 train/test ratio with no crossover of participants between the training and test set, following the same method as [30]. The hyperparameters are varied within these ranges: $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 25\}$, $\alpha \in \{0.1, 0.01, 0.001\}$, and $\beta \in \{0.1, 0.01, 0.001\}$.

2.3 Phenotype validation

We validate the phenotypes in various ways, to ensure that they are appropriate and suitable for the intended tasks. We first examine the learned phenotype model to understand how it has characterized health status. We look at the vocabulary, visualized by heatmaps and wordclouds. We use these to make sense of what types of insights the mixed-membership model has learned and how it has characterized health status, e.g., has it learned differences in types of illness experience (gi-based vs pain-based) or has it learned differences based on severity of the illness experience that week. We compare the learned phenotype model to a baseline phenotype model (constructed using rules based on the 'How was your day?' Phendo question) in various ways.

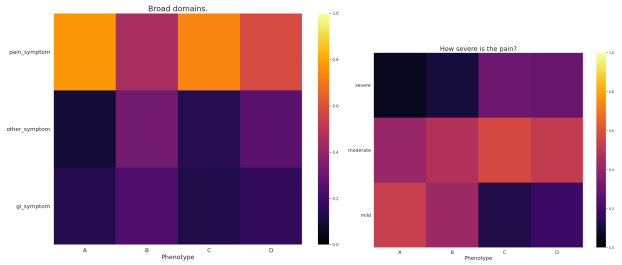
2.4 Phenotype evaluation

We evaluate the phenotypes (referred to as "AI-generated health statuses" in the user study) by consulting individuals with endometriosis to assess whether the phenotypes represent meaningful facets of illness experience and the potential real-world applicability. The user evaluation consists of two parts — the first to evaluate phenotypes with users' real-world data to see if the health statuses align with their understanding of their illness; the second to evaluate the temporal component of the health statuses to see if they help individuals evaluate changes in health status over time. Participants are asked to think-aloud as they complete each task. Along with this, participants are asked to assess the difficulty and certainty, along with their agreement of the AI-generated health status. We recruited N = 5 individuals who previously used the Phendo app to track their experience of illness (minimum 6 weeks of data), to evaluate the phenotypes.

From other work [23], individuals are more interested in when they will have flare-ups, rather than when they will have "good" days; they also prefer false positives (predicting a flare-up that does not happen) over false negatives (predicting no flare-up when in reality symptoms do worsen).

3 Results

Final model. We learned a model with $K = 4$ topics and hyperparameters $\alpha = 0.01$ and $\beta = 0.1$. The mixed-membership model learned a largely severity-based representation of health statuses. This was not a given, and the model could have learned various other dynamics instead (e.g., based on body system or type of management that was helpful). This severity-based characterization was consistent across all K topics and α/β hyperparameters. This robustness provides confidence in the model, and that a severity-based phenotype is the most appropriate approach. It also helps

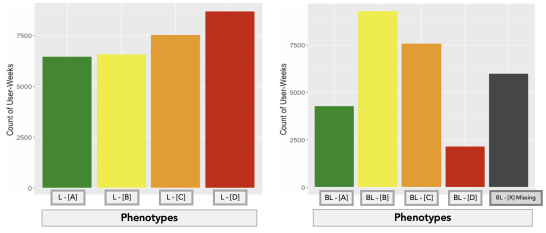


(a) 'What symptoms are you experiencing?' (top-*pain*; middle-*other*; bottom-*GI*) **(b) 'How severe is the pain?' (top-*severe*; middle-*moderate*; bottom-*mild*)**

Figure 1: Learned Phenotype Model: Symptoms and Severity

us to learn about the underlying population and experience of illness. Heatmaps of *symptoms* and *pain severity* for the final learned phenotype model are shown in Fig 1.

Validation. In the learned model, each of the four phenotypes is well-represented, with some variation. The baseline model, on the other hand, is more skewed towards some phenotypes with less representation of others (i.e., the best and worst phenotypes have fewer assignments in the dataset). The baseline model also has a lot of "missing" assignments that are not missing in the learned model, due to the "How was your day?" question not being answered. The distributions are shown in Fig 2. While there is a slight correlation between the learned and baseline phenotypes, the learned model does not neatly align with the "How was your day?" variable. Thus, the learned phenotypes are useful to characterize what is going on with individuals and their health, providing richness and nuance.



(a) Learned Phenotypes **(b) Baseline Phenotypes**

Figure 2: Phenotype Model: Distribution of Phenotypes

We also created user timelines of all user-weeks from an individual, shown in Fig 3. We find that individual users are not assigned to a single phenotype across their entire timeline (i.e., the model is not learning user-level dynamics), and that there is variation in phenotype assignments. There is wide heterogeneity across each individual's timeline and across different users. These different dynamics can help us to learn about the experience of disease, giving insights about our population. We also examine weeks where users menstruate and do not find an association across phenotypes. This gives us further insights into the population of users — bad weeks are not only during menstrual periods.

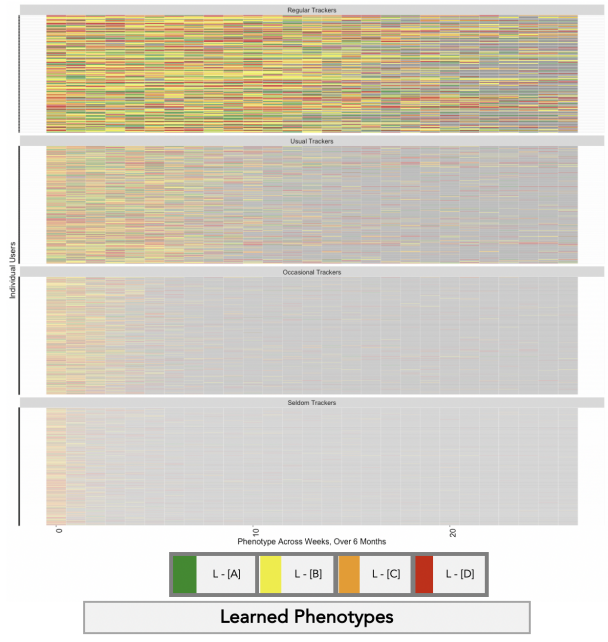
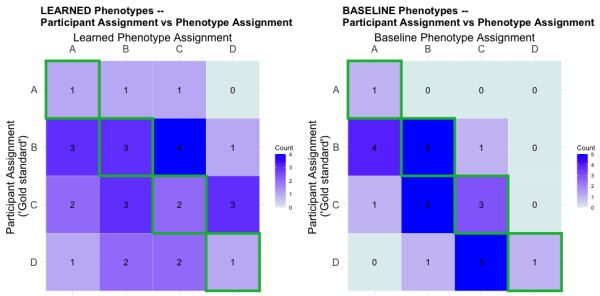


Figure 3: Learned Phenotype Model: Temporal Plot of Phenotypes Across Engagement Levels

Evaluation. There is limited alignment between user assignments and the phenotypes — participants' assignments matched the learned phenotypes 23% of the time (n = 7), and matched the baseline phenotypes 33% of the time (n = 10). The baseline phenotypes tended to under-estimate individual's health status (i.e., among disagreements, in 16 cases, the AI-generated health status was better than the user's rating, while in only 1 case was it worse). While the learned phenotypes sometimes under-estimated the severity of individuals' health statuses, incorrect assignments made by the learned model were more likely to be more severe than less severe, when compared to the users' own assessments of their health status (i.e., in 13 cases the AI-generated health status was better than the user's rating, while in 10 cases the AI-generated health status was worse than the user's rated health status). This information is visualized in Fig 4. This aligns with the user's preference for the AI to err on the side of assigning a status that is worse than reality, rather than assigning one that is better than reality.



(a) Heatmap of the learned phenotype assignments vs the user assignments **(b) Heatmap of the baseline phenotype assignments vs the user assignments**

Figure 4: AI-generated health status vs user assignments (for the primary assessment), across the learned and baseline phenotypes. The "matched" boxes (where both the phenotype assignment and the user assignment match) are outlined in green.

Results from across the two repeated assignments where we showed participants the same data suggest limited test-retest reliability. Across all users and instances, individuals made inconsistent assignments 33% of the time. Despite their lack of consistency, participants rated themselves as relatively certain and felt the task was somewhat easy. At the same time, very few participants changed their assignments after finding out what the AI-generated health assignment was, despite the frequent discrepancy between their assignment and that of the phenotype model. Individuals kept the

same assignment in 93% of cases (n = 56), and changed their assignments in only 7% of cases (n = 4).

To contextualize these findings, we turn to the insights from the *qualitative data* captured during the study visits. Participants talked about using a combination across a range of indicators to assess their health status. This mirrors what the learned phenotype relied on to make the assignment, more than the baseline phenotype, which used only the single Day Rating indicator. Various aspects were described as impacting the difficulty of making health assessments. Weeks with very little data were hard for participants to make assessments. They also had a hard time when they felt that different days of the week would give different assessments, or may sway a week's health status towards better or worse than their overall assessment may otherwise be. Individuals also speculated that they tend to minimize their pain and do not like to log that their experience was too bad. Participants also talked about making sense of disagreements between their assignment and the phenotype. While they explained that they could sometimes see why the AI made a particular assignment, they largely relied on their own assessments over that of the AI. Participants frequently emphasized the aspects of their data that strongly suggested to them the health status that they assigned, e.g., pain medications or severe pain. They were also frustrated when the AI over-estimated their health status.

Participants were largely able to provide accurate assessments of health status over time for all of the cases (i.e., surgery, self-management, clinical summary), even without access to the AI-generated health statuses. But they had some uncertainty about it and it took a long time for them to review the detailed self-tracked data. Some of the cases were more challenging than others to make evaluations. For the clinical summary case, participants said the AI-generated health status gave them confidence in their own evaluations. Participants were all optimistic about using such a technology if it were available to them. They felt it would help them in assessing their own health status, although they viewed themselves as the experts and would want control over the AI's behavior and outputs. Participants also described potential value in taking the health status reports to their healthcare providers.

4 Discussion

Results from the user study highlight the nuanced relationship between participants' self-assessments of health status and health status phenotypes, offering insights into their thought processes and where the current phenotypes might be improved. Participants demonstrated a clear reliance on a wide range of indicators, including symptom severity, medication usage, and activities of daily living, to make health assessments. These indicators, while overlapping among participants, were often interpreted and weighted differently, reflecting the individualized and subjective nature of self-assessment. Interestingly, while participants valued the AI-generated health statuses as a reference point, they overwhelmingly maintained their own assessments, showcasing their role as the primary experts of their lived experiences. At the same time, there were significant discrepancies in repeated evaluations of the same data. This requires further study to fully understand and reconcile with a machine-readable health status phenotyping model. Despite this, participants recognized the potential utility of AI-enabled tools in validating their experiences and assisting with summarization, particularly to be used as a resource for communicating with their care teams. Additionally, while the AI-generated health statuses enhanced participants' efficiency in assessing longitudinal health trends, they did not necessarily increase their confidence or make decision-making about their evaluations easier.

In line with human-centered AI, these findings underscore the importance of designing intelligent systems that complement rather than override user expertise, emphasize transparency in AI reasoning, and account for the complexity of individual health narratives. These findings also call for innovation in how users can remain "in the loop" with these models, or otherwise enhance autonomy and control over how their data represent their experiences. Integrating AI as a supportive, participatory tool has the potential to enhance both self-awareness and patient-provider interactions, but first further work is required so that the outputs more closely align with users' expectations and lived realities.

The evaluation of the phenotypes also gave insight into how the phenotypes could be improved. In the user study, many participants talked about using medication to make their assessments. This information was not fully incorporated into model training. In the current model, medication was included as a simple binary (took medication, yes), and future work could map and categorize the user-entered medications to identify pain medications. This is an insight directly garnered from individuals in the user study. Future work on the phenotypes could include improving how current data are used (e.g., mapping pain medications), using more advanced ML techniques to use other existing data (e.g., NLP methods to use the open-ended journal text), or incorporating additional datatypes (e.g., passive sensing).

5 Conclusion

This work showcases an example of research at the intersection of Health and HCI that is important, compelling, and challenging. While we offer sone solutions and directions for development, we also raise new questions and highlight real-world challenges that require further exploration. Through this workshop, we hope to engage with other researchers who are tackling similar problems.

Acknowledgments

We thank the participants of this research. This research was supported by the National Library of Medicine (T15LM007079 and R01LM013043)

References

[1] P Acién and I Velasco. 2013. Endometriosis: A Disease That Remains Enigmatic. *Int Sch Res Notices* (2013).
[2] SK Agarwal, C Chapron, LC Giudice, MR Laufer, N Leyland, SA Missmer, SS Singh, and HS Taylor. 2019. Clinical diagnosis of endometriosis: a call to action. *Am J Obstet Gynecol* (2019).
[3] D Buls and J Rooksby. 2017. Technology for Self-Management of Rosacea: A Survey and Field Trial. In *Proc ACM CHI Conf* (NYC, NY, USA) *(CHI EA '17)*.
[4] J Corbin and A Strauss. 1985. Managing chronic illness at home: Three lines of work. *Qual Sociol* (1985).
[5] EAF Dancet, S Apers, JAM Kremer, WLDM Nelen, W Sermeus, and TM D'Hooghe. 2014. The Patient-Centeredness of Endometriosis Care and Targets for Improvement: A Systematic Review. *Gynecol Obstet Invest* 78, 2 (2014), 69–80.
[6] G Demiris and L Kneale. 2015. Informatics Systems and Tools to Facilitate Patient-centered Care Coordination. *Yearb Med Inform* 24, 1 (2015), 15–21.

[7] N Elhadad. 2021. Phendo app available at Apple's App store. https://itunes.apple.com/us/app/phendo/id1145512423.

[8] N Elhadad. 2021. Phendo app available at Google Play. https://play.google.com/store/apps/details?id=com.appliedinformaticsinc.phendo.

[9] DA Epstein, C Caldeira, MC Figueiredo, X Lu, LM. Silva, L Williams, JH Lee, Q Li, S Ahuja, Q Chen, P Dowlatyari, C Hilby, S Sultana, EV Eikey, and Y Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2020).

[10] MC Figueiredo and Y Chen. 2020. Patient-Generated Health Data: Dimensions, Challenges, and Open Questions. *Foundations and Trends® in Human–Computer Interaction* (2020).

[11] EI Geukens, S Apers, C Meuleman, TM D'Hooghe, and EAF Dancet. 2018. Patient-centeredness and endometriosis: Definition, measurement, and current status. *Best Practice & Research Clinical Obstetrics & Gynaecology* (2018).

[12] S Grimme, SM Spoerl, S Boll, and M Koelle. 2024. My Data, My Choice, My Insights: Women's Requirements when Collecting, Interpreting and Sharing their Personal Health Data. In *Proc ACM CHI Conf* (New York, NY, USA) *(CHI '24)*.

[13] C Hajat and E Stein. 2018. The global burden of multiple chronic conditions: A narrative review. *Preventive Medicine Reports* (2018).

[14] R Karkar, J Zia, J Schroeder, DA Epstein, LR Pina, J Scofield, J Fogarty, JA Kientz, SA Munson, and R Vilardaga. 2017. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. In *Proc ACM CHI Conf* (Denver, CO, USA). ACM Press.

[15] B Klijs, WJ. Nusselder, CW. Looman, and JP Mackenbach. 2011. Contribution of Chronic Disease to the Burden of Disability. *PLOS ONE* (2011).

[16] JD Lee and A Hohler. 2014. Communication challenges in complex medical environments. *Continuum (Minneapolis, Minn.)* (2014).

[17] C. G. Nicholas Mascie-Taylor and Enamul Karim. 2003. The Burden of Chronic Disease. *Science* (2003).

[18] M McKillop, L Mamykina, and N Elhadad. 2018. Designing in the dark: eliciting self-tracking dimensions for understanding enigmatic disease. In *Proc ACM CHI Conf.*

[19] M McKillop, N Voigt, R Schnall, and N Elhadad. 2016. Exploring self-tracking as a participatory research activity among women with endometriosis. *J Particip Med* (2016). Issue e17.

[20] RV Milani and CJ Lavie. 2015. Health Care 2020: Reengineering Health Care Delivery to Combat Chronic Disease. *Am J Med* (2015).

[21] A Ostropolets, RJ Chen, L Zhang, and G Hripcsak. 2020. Characterizing physicians' information needs related to a gap in knowledge unmet by current evidence. *JAMIA Open* 2 (2020).

[22] CJ Peek, MA Baird, and E Coleman. 2009. Primary care for patient complexity, not only disease. *Families, Systems, & Health* 27 (2009).

[23] A Pichon, J Blumberg, L Mamykina, and N Elhadad. 2025. The Voice of Endo: Leveraging Speech for an Intelligent System That Can Forecast Illness Flare-ups. In *Accepted for publication in: Proc ACM CHI Conf.*

[24] A Pichon, K Schiffer, E Horan, B Massey, S Bakken, L Mamykina, and N Elhadad. 2020. Divided We Stand: The Collaborative Work of Patients and Providers in an Enigmatic Chronic Disease. *Proc ACM CSCW Conf* (2020).

[25] F Rajabiyazdi, C Perin, J Babione, M Santana, J Kaufman, W Ghali, P Sargious, S Carpendale, and J Tropiano. 2016. Involving Patients in their Care Plan: Patients' and Care providers' Perspectives. In *Proceedings of the CHI Workshop on Interactive Systems in Healthcare (WISH'16)*.

[26] J Schroeder, CF Chung, DA Epstein, R Karkar, A Parsons, N Murinova, J Fogarty, and SA Munson. 2018. Examining Self-Tracking by People with Migraine: Goals, Needs, and Opportunities in a Chronic Health Condition. In *Proceedings of the 2018 Designing Interactive Systems Conference* (New York, NY, USA) *(DIS '18)*.

[27] J Schroeder, J Hoffswell, CF Chung, J Fogarty, S Munson, and J Zia. 2017. Supporting Patient-Provider Collaboration to Identify Individual Triggers using Food and Symptom Journals. In *Proc ACM CSCW Conf* (Portland, OR, USA) *(CSCW '17)*.

[28] A Strauss, S Fagerhaugh, B Suczek, and C Wiener. 1982. Sentimental work in the technologized hospital. *Sociol Health Illn* 4, 3 (1982), 254–278.

[29] FC Udegbe, OR Ebulue, CC Ebulue, and CS Ekesiobi. 2024. The role of artificial intelligence in healthcare: a systematic review of applications and challenges. *International Medical Science Research Journal* (2024).

[30] I Urteaga, M McKillop, and N Elhadad. 2020. Learning endometriosis phenotypes from patient-generated data. *npj Digital Medicine* (2020).

[31] I Urteaga, M McKillop, S Lipsky-Gorman, and N Elhadad. 2018. Phenotyping endometriosis through mixed membership models of self-tracking data. In *Proc Machine Learning for Health Care (MLHC '18)*.

[32] EH Wagner, BT Austin, and M Von Korff. 1996. Organizing Care for Patients with Chronic Illness. *The Milbank Quarterly* (1996).

[33] D Yach, C Hawkes, CL Gould, and KJ. Hofman. 2004. The Global Burden of Chronic DiseasesOvercoming Impediments to Prevention and Control. *JAMA* (2004).

[34] KT Zondervan, CM Becker, K Koga, SA Missmer, RN Taylor, and P Viganò. 2018. Endometriosis (Primer). *Nature Reviews: Disease Primers* (2018).