Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SymptomAI: Toward a Conversational AI Agent for Everyday Symptom Assessment

Daniel McDuff

dmcduff@google.com

Google https://orcid.org/0000-0001-7313-0082

Joseph Breda

Google

Fadi Yousif

Google

Beszel Hawkins

Google Research

Marinela Cotoi

Google

Miao Liu

Google

Ray Luo

Google

Po-Hsuan Chen

Google https://orcid.org/0000-0002-0083-4991

Mike Schaekermann

Google https://orcid.org/0000-0002-1735-9680

Samuel Schmidgall

Google DeepMind

Xin Liu

Google Research https://orcid.org/0000-0002-9279-5386

Girish Narayanswamy

Google

Samuel Solomon

Google

Maxwell Xu

Google

Xiaoran Fan

Google Research https://orcid.org/0000-0002-6368-9250

Longfei Shangguan

Google

Anran Wang
Google Research

Bhavna Daryani
Google

Buddy Herkenham
Google

Cara Tan
Google

Mark Malhotra
Google Research

Shwetak Patel
Google

John Hernandez
Google Research

Quang Duong
Google

Yun Liu
Google https://orcid.org/0000-0003-4079-8275

Zach Wasson
Google

Dimitrios Antos
Google

Bob Lou
Google https://orcid.org/0000-0003-4723-5416

Matthew Thompson
Google

Jonathan Richina
Google

Anupam Pathak
Google

Nichole Young-Lin
Google

Jake Sunshine
Google https://orcid.org/0000-0001-9133-6922

Biological Sciences - Article

Keywords:

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9684240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. The authors receive Google equity as part of their typical compensation.

SymptomAI: Toward a Conversational AI Agent for Everyday Symptom Assessment

Joseph Breda[1,†], Fadi Yousif[1], Beszel Hawkins[1], Marinela Cotoi[1], Miao Liu[1], Ray Luo[1], Po-Hsuan Cameron Chen[1], Mike Schaekermann[1], Samuel Schmidgall[2], Xin Liu[1], Girish Narayanswamy[1], Samuel Solomon[1], Maxwell A. Xu[1], Xiaoran Fan[1], Longfei Shangguan[1], Anran Wang[1], Bhavna Daryani[1], Buddy Herkenham[1], Cara Tan[1], Mark Malhotra[1], Shwetak Patel[1], John B. Hernandez[1], Quang Duong[1], Yun Liu[1], Zach Wasson[1], Dimitrios Antos[1], Bob Lou[1], Matthew Thompson[1], Jonathan Richina[1], Anupam Pathak[1], Nichole Young-Lin[1], Jake Sunshine[1,‡] and Daniel McDuff[1,‡]

[‡]Equal Leadership, [1]Google Research, [2]Google DeepMind, [†]Work done while at Google Research

Language models excel at diagnostic assessments on curated medical case-studies and vignettes, performing on par with, or better than, clinical professionals. However, existing studies focus on complex scenarios with rich context making it difficult to draw conclusions about how these systems perform for patients reporting symptoms in everyday life. We deployed *SymptomAI*, a set of conversational AI agents for end-to-end patient interviewing and differential diagnosis (DDx), via the Fitbit app in a study that randomized participants ($N = 13,917$) to interact with five AI agents. This corpus captures diverse communication and a realistic distribution of illnesses from a real world population. A subset of 1,228 participants reported a clinician-provided diagnosis, and 517 of these were further evaluated by a panel of clinicians during over 250 hours of annotation. *SymptomAI* DDx were significantly more accurate ($OR = 2.56$, $p < 0.001$) than those from independent clinicians given the same dialogue in a blinded randomized comparison. Moreover, agentic strategies which conduct a dedicated symptom interview that elicit additional symptom information *before* providing a diagnosis, perform substantially better than baseline, user-guided conversations ($p < 0.001$). An auxiliary analysis on 1,509 conversations from a general US population panel validated that these results generalize beyond wearable device users. We used *SymptomAI* diagnoses as labels for all 13,917 participants to analyze over 500,000 days of wearable metrics across nearly 400 unique conditions. We identified strong associations between acute infections and physiological shifts (e.g., $OR > 7$ for influenza). While limited by self-reported ground truth, these results demonstrate the benefits of a dedicated and complete symptom interview compared to a user-guided symptom discussion, which is the default of most consumer LLMs.

1. Introduction

Consumer health information seeking patterns have undergone a global transformation in the 21st century with the rise of the Internet (Jia et al., 2021) and, more recently, the introduction of large language models (LMs) (Gallup). With up to one in five conversational AI queries relating to medical knowledge (Sumner et al., 2025), and millions of people using it for medical advice regularly (Shahsavar et al., 2023), AI is on track to becoming a primary interface that people approach for medical information needs (Ayers et al., 2023). Infact, a recent investigation into the types of personal guidance sought through one conversational AI platform found health and wellness to be the most popular topic, representing over a quarter of guidance-seeking conversations (Shen et al., 2026).

Close to 20% of health-related AI chat conversations involve symptom assessment or condition discussion (Costa-Gomes et al., 2026). This trend toward self-guided medical assessment through technology predates conversational AI. Increases in search engine queries for symptoms predict decreases in outpatient visits for the same medical concerns (Heumann and Steinhubl, 2025). Consequently, a variety of online symptom checkers have emerged that enable patients to retrieve a set of possible

Corresponding author(s): joebreda, jakesunshine, dmcduff@google.com
© 2026 Google. All rights reserved

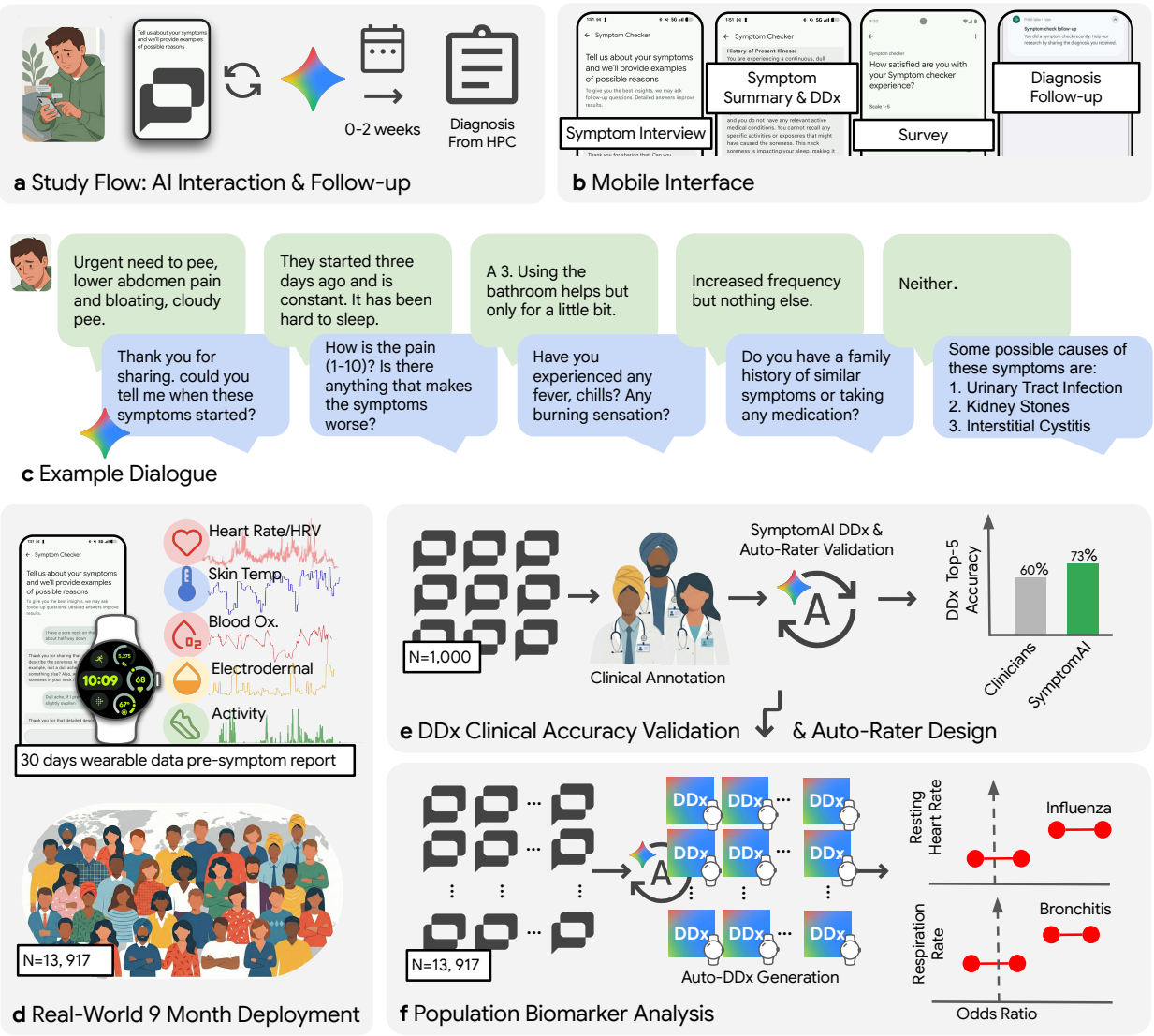


Figure 1 | ***SymptomAI* Study.** (a-b) Experimental deployment study procedure of *SymptomAI* for end-to-end patient interviewing and generative AI differential diagnosis (DDx) for symptom assessment that were benchmarked against study participant-reported diagnoses recieved from a Health Care Provider (HCP). (c-d) This led to a large dataset (N=13,917) of naturalistic symptom conversations communicated by laypeople paired with recent wearable data. (e) We leveraged clinical expert annotation to validate *SymptomAI* DDx against and to inform the development of an LLM verifier (i.e., auto-rater) for expanding validation beyond the clinical evaluation sub-sample. (f) Leveraging *SymptomAI* as a phenotype labeler enables phenome-wide analysis of biosignals across the study population.

candidate diagnoses from a set of self-reported symptoms; however, traditional solutions are extremely limited with diagnostic accuracies ranging between 20-40% (Gilbert et al., 2023; Semigran et al., 2015; Wallace et al., 2022). This is significant as these initial symptom assessments often serve as a primary entry point for downstream medical care.

Clinical history-taking (i.e., natural language exchange) alone is estimated to provide the basis for 75-80% of diagnoses (Hampton et al., 1975; Peterson et al., 1992; Roshan and Rao, 2000), representing a

significant opportunity for impact through accurate diagnostic LMs. Harnessing this opportunity could have a significant impact on public health as access to clinical expertise is both episodic and globally scarce (Crisp and Chen, 2014). More broadly, improving access to medical reasoning expertise could positively impact the quality, accessibility, consistency, and affordability of medical intervention, both through empowering self-initiated assessment and through supplementing professional face-to-face guidance when available.

Towards this, LMs have demonstrated performance on par with, or better than, clinical professionals on curated medical knowledge benchmarks (Arora et al., 2025; Bedi et al., 2025; Nori et al., 2023; Saab et al., 2024) and archival medical case studies (Kanjee et al., 2023; Manrai et al., 2026; McDuff et al., 2025). However, these text corpora used for evaluation have been limited to synthetic examples (Arora et al., 2025; Liu et al., 2026), single-turn medical question-answering (Nori et al., 2023; Saab et al., 2024; Singhal et al., 2023), or highly detailed reports of atypical, challenging medical scenarios (McDuff et al., 2025), none of which are truly representative of the type of information communicated conversationally by a patient in an everyday interaction through a digital interface nor the distribution of reported symptoms and illnesses across the population. Analysis of patient interviews simulated by trained patient actors has demonstrated improved performance in more realistic multi-turn scenarios, showing the potential of conversational AI for history-taking (Tu et al., 2025). While related work has explored conversational AI for specialized care (O'Sullivan et al., 2026; Palepu et al., 2025a), disease management (Palepu et al., 2025b), and multimodal reasoning (Saab et al., 2025), as well as frameworks for physician oversight (Vedadi et al., 2025), applying these systems to general symptom assessment by laypeople introduces unique challenges. An evaluation of diagnosis through human-AI interaction revealed that the involvement of laypeople in communicating necessary context significantly degraded AI diagnostic accuracy compared to AI applied directly to clinical vignettes (from 94.5% to 34.5%) (Bean et al., 2025; Goh et al., 2024). This decrease in performance is largely due to the incomplete or misrepresented information provided by non-experts, indicating the criticality of testing diagnostic AI, when possible, with real users in naturalistic conditions. While recent work has begun to assess conversational AI involving laypeople with real health needs on subjective measures like perceived helpfulness in modest-scale studies (Sayres et al., 2026), a large-scale evaluation of conversational AI for layperson symptom assessment using clinical measures like diagnostic accuracy has not yet been demonstrated.

To comprehensively understand the performance of conversational AI for providing accessible medical information to the broader population, in real-world contexts, it is important to evaluate its performance through an integrated study involving people with real health needs, assessing the ability of a general-purpose symptom checker to: (1) conduct flexible and personalized patient interviews to elicit appropriate context, (2) produce accurate DDx given context provided by general users, and (3) maintain performance across a range of real-world conditions.

We conducted a strictly experimental research study of *SymptomAI*, a conversational AI agent, built on top of Gemini. This study-specific system was operationalized through the Fitbit Labs research environment in the Fitbit mobile application* from June 2025 to April 2026. The system was designed to explore the feasibility of both guiding participants through a series of symptom-related questions and providing a set of possible associated reasons with relevant educational information to research study participants (see Figure 1). The study was undertaken under informed consent (Advarra, Maryland USA: GH-SCD-001). Our investigation resulted in 13,917 multi-turn conversations in which research study participants voluntarily described their health symptoms to *SymptomAI*. During this investigation, we randomized the participants across five study arms, representing different agent prompting strategies ranging from highly structured history of present illness (HPI) interviews

*https://play.google.com/store/apps/details?id=com.fitbit.FitbitMobile&hl=en_US

based on canonical medical history taking questions (e.g., onset, location, symptom characterization, symptom provocation/palliation, symptom quality, symptom severity and associated symptoms) to a fully dynamic conversational agent, specifically a variant of the 'Wayfinding AI' system from (Sayres et al., 2026). Along with these conversations, we asked research study participants to report any diagnoses they received from healthcare provider interactions at the outset of, and two weeks after, their engagement with *SymptomAI* through an in-app survey. To further ground our evaluation, we conducted a human-expert annotation study where a panel of three board-certified Family Medicine physicians with >35 years of post-residency experience across primary care, urgent care, and academic medicine provided an independent DDx based on AI conversation transcripts. Each clinician then assessed the top-5 accuracy of DDx provided by both real clinicians and by *SymptomAI* side by side, while blinded to the author of the DDx and the participant's actual diagnosis, for a subset of conversations. Clinicians were also requested to provide quality assessments of all study materials (conversation chat logs, participant-reported diagnoses, and all DDx lists). The exploratory associations generated by the system were not shared with the participants' actual primary care providers and had no bearing on their clinical treatment plan.

Building on the clinical expert validation, we employed an auto-rater to assess the DDx accuracy of *SymptomAI* across the entire cohort of participants who self-reported a diagnosis (N=1,228). To then leverage our full study cohort (N=13,917), we treated the top diagnosis generated by *SymptomAI* as a reference label for these research study participants, enabling a phenome-wide association study (PheWAS) (Bastarache et al., 2022) of nearly 400 unique medical diagnoses with over 500,000 days of wearable biosignals. This would have been infeasible to perform at comparable scale without the use of AI-generated reference diagnoses. We show that AI generated diagnoses from *SymptomAI* – particularly those of acute respiratory infections – share trends with wearable biosignals, potentially enabling future research analysis of wearable biosignals for predicting symptom onset. Such predictions could be used to trigger future *SymptomAI* conversations amongst users suspected of respiratory (and potentially non-respiratory) infectious diseases.

2. Results

2.1. Conversations That Elicit More Information Outperform User-Guided Conversations

Participants were randomly assigned to one of five study arms, each employing a different prompting strategy. Prompting strategies that explicitly elicited more information from the user through follow-up questions outperformed the base user-guided condition (Chi-Squared Test - $p < 0.001$).

Arm 1: Base was only instructed to restrict responses to medical and health topics, reflecting the base performance of Gemini 2.0 Flash without specialized prompting to guide the conversation. *Arm 2: Fixed canonical questions (Fixed Canonical) & Arm 3: Flexible canonical questions (Flexible Canonical)* were both based on canonical medical history taking questions, representing "low agency" condition where *SymptomAI* was instructed to follow a structured interview procedure with a set of prescriptive questions. *Arm 2* asked a fixed set of prescriptive questions irrespective of the users responses, while *Arm 3* was allowed flexibility to drop irrelevant questions during the interview. *Arm 4: Dynamic with live updates (Dynamic Live) & Arm 5: Dynamic with final output only (Dynamic Final)* gave *SymptomAI* full agency over which follow up questions to ask and only restricted the number of turns before making a final DDx. *Arm 4* provided intermediate best-effort DDx at every turn, while *Arm 5* provided only a final DDx at the end of the conversation. For a full description of the prompting strategies, see Appendix E.

Arms 2-5 explicitly elicited more information from the user, either through prescribed follow-up questions or enforcing multi-turn conversations of a minimum number of turns. This general strategy, a model-guided interview that elicits more information, resulted in an average of 27.57% higher

accuracy than the base user-guided prompting strategy which did not explicitly elicit more information and relies on users to provide or not provide information as part of the symptom checking experience. Specifically, each experimental arm independently outperformed the base condition via Chi-Squared (*arm 2*; $p = 0.003$, *arm 3*; $p < 0.001$, *arm 4*; $p = 0.003$, *arm 5*; $p = 0.022$). Moreover, prompting strategies that did not explicitly pre-define questions (*arms 4 & 5*; combined accuracy 72.07%) yielded comparable performance to strategies utilizing canonical, clinician-defined history-taking questions (*arms 2 & 3*; combined accuracy 77.51%), with no statistically significant difference observed via Chi-Squared test ($p = 0.235$). Figure 2c shows the accuracy stratified by arm, while Figure 2d shows the distribution of total user words provided in the interview. This speaks to *SymptomAI*'s ability to conduct effective patient history of present illness interviews autonomously based on natural conversational trajectory, yielding accuracy results similar to those obtained through asking common medical history questions. Conversely, the results also demonstrate that models adopting canonical history questions taught in medical training perform similarly to models not specifically constrained in what they can ask but tasked with arriving at an accurate differential diagnosis.

2.2. Clinical Experts Prefer *SymptomAI* Over Clinician-Generated Differential Diagnoses

A cohort of 517 participants were selected for clinical evaluation according to the method described in Section A.4. All clinical evaluations were conducted as retrospective reviews of the conversation transcripts by expert raters who did not engage in any direct interaction or clinical care with the study participants. For each case, we generated three DDx per case, one from *SymptomAI* and two from independent clinicians ("baseline clinicians") reviewing the conversation transcripts. To ensure unbiased evaluations, another independent clinician ("clinical rater") ranked these DDx while blinded to both the DDx authors and the ground-truth diagnoses, relying exclusively on their clinical judgment.

The clinical raters ranked the *SymptomAI* DDx as the best option in over 50% of the cases, indicating a significant preference over chance (odds ratio of 2.34, one-sided binomial test against expected $P = 0.33$, $n = 517$, $p < 0.001$). This is also supported by a Cohen's h of 0.42, indicating a clear, consistent, small-to-medium effect size, in favor of *SymptomAI*. Figure 2a shows the proportion of 1st, 2nd and 3rd DDx across *SymptomAI* and baseline clinicians rated by the clinical raters. Importantly, the clinical raters' preference for *SymptomAI* is most pronounced in the subset of conversations clinician raters deem highest quality (see Supplemental Figure 11). Because these highest-quality conversations provide the most complete clinical context, they represent the most rigorous baseline for comparison of *SymptomAI* DDx against baseline clinicians with similar context.

2.3. *SymptomAI* DDx is More Accurate than Clinicians

To quantitatively assess the accuracy of DDx, on the cohort of 517 cases, the clinical raters reviewed the DDx alongside the ground truth diagnosis, while blinded to the DDx author.

SymptomAI demonstrated higher top-5 DDx accuracy over the baseline clinician's DDx (McNemar's Test: Median OR = 2.56, 95% CI Cohen's g [0.18, 0.26], $p < 0.001$). Figure 2b shows the average top-5 accuracy assigned by the clinical raters for *SymptomAI* and the baseline clinicians. Importantly, these results hold for the subset of conversations rated as highest quality by the clinical raters. Because these conversations provide complete clinical information, they offer the most rigorous and ideal baseline for comparison where the baseline clinicians have the full context for a complete DDx. Figure 2e shows the top-5 accuracy of *SymptomAI* and baseline clinicians' DDx stratified across conversation quality, defined as whether the conversation contains sufficient context to produce an accurate DDx as assessed by the clinical raters.

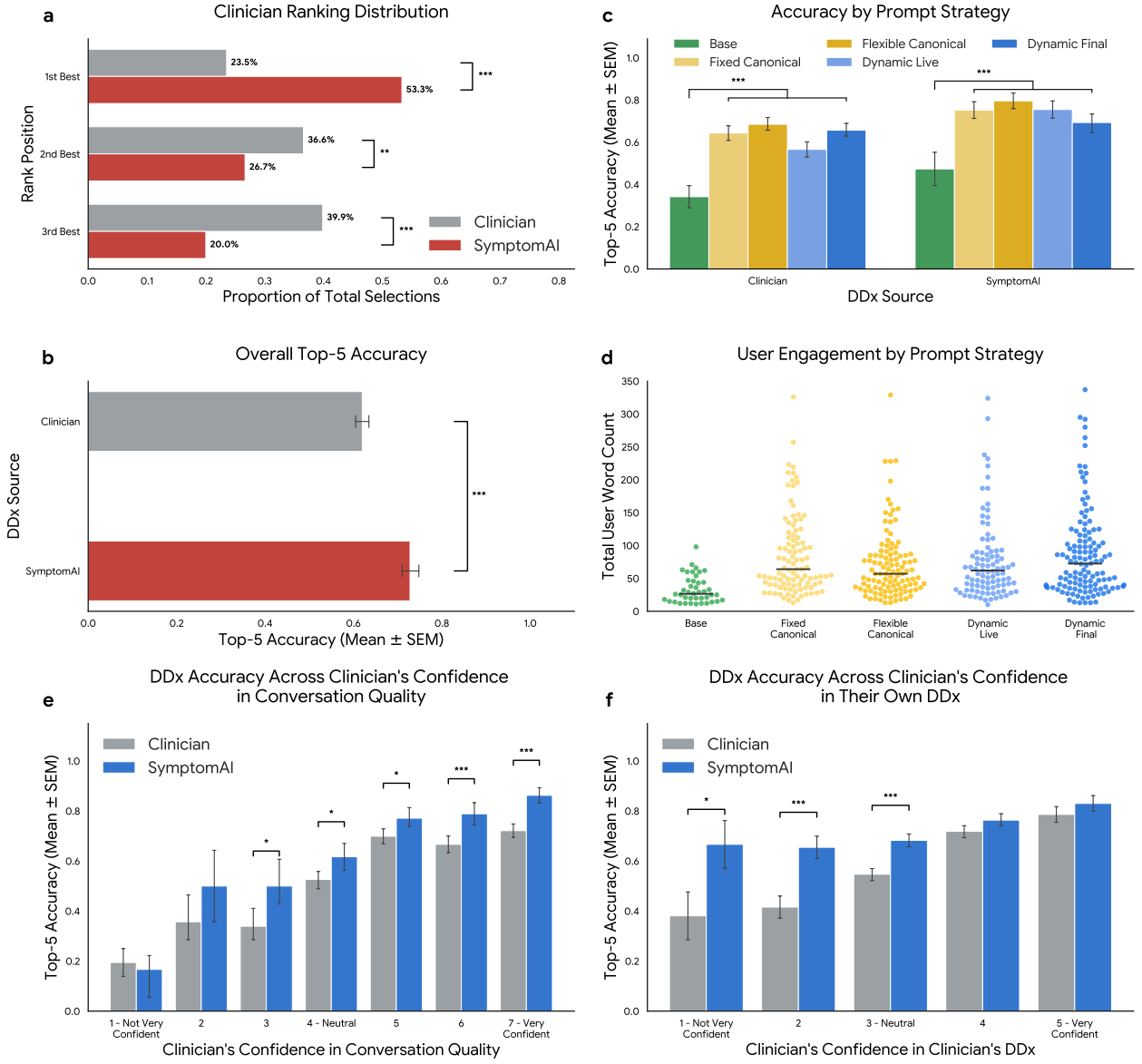


Figure 2 | Clinical evaluation and user engagement of *SymptomAI*. (a) Proportion of *SymptomAI* and clinician DDx (normalized by category total) ranked by blinded clinicians as 1st, 2nd, and 3rd position amongst a randomized list of 3 possible DDx lists for each conversation (one *SymptomAI* and two clinician baselines per trial). (b) The average top-5 accuracy assigned by clinicians to DDx produced by clinicians and *SymptomAI*. (c) Top-5 accuracy of clinicians and *SymptomAI* stratified by conversation prompting strategy. (d) Total user words sent across all user messages across each prompting strategy. Horizontal line denotes median word count. (e) The top-5 accuracy assigned by clinicians to *SymptomAI* and baseline clinician-generated DDx stratified by clinician's confidence that the conversation contained enough information to support a plausibly accurate DDx. (f) The top-5 accuracy assigned by clinicians to *SymptomAI* and baseline clinician-generated DDx stratified by clinician's rating of confidence in their own DDx for that conversation.

2.4. Robustness of *SymptomAI* on Low Information Conversations

Figure 2f shows the top-5 accuracy for *SymptomAI* and clinician DDx for conversations stratified by the clinician's confidence in their own DDx. While clinicians and *SymptomAI* performed similarly well on conversations where the clinicians felt confident in their own DDx, *SymptomAI* significantly outperformed the clinicians for the conversations where the clinicians felt neutral or not confident in their own DDx via McNemar's test ($p < 0.001$). This indicates that *SymptomAI* is robust to factors that degrade human expert DDx performance and confidence.

2.5. Representativeness of the Study Population

The Clinical Evaluation Sub-Sample is Representative of the Full Study Population. The sample of conversations used in the clinical evaluation was randomly sampled from the subset of the study population which had self-reported a diagnosis obtained from a healthcare provider. We assessed the representativeness of this clinical evaluation subsample relative to the total study population and found low effect size, and therefore no practically meaningful distributional shifts in demographic covariates: age $D = 0.060$ ($p = 0.001$), gender $V = 0.058$ ($\chi^2(3) = 50.23$, $p < 0.001$), and weight $D = 0.026$ ($p = 0.490$). This suggests that the subsample of research study participants who self-reported a diagnosis did not introduce detectable selection bias across demographics. We find only a small shift in diagnosis category Phecode: $V = 0.109$ ($\chi^2(390) = 553.73$, $p < 0.001$). A model stress test predicting whether participants reported a diagnosis resulted in poor discriminative ability ($AUC = 0.591$), suggesting further, through multivariate control, that the propensity to self-report was not strongly associated with user demographics or the category of illness the research study participants faced.

***SymptomAI* Top-5 Accuracy Remains Consistent with the General Population.** Sampling research study participants via Fitbit Labs in the Fitbit app raises questions about whether reported symptoms and DDx accuracy is similar to those expected from a general cross-section of the US population. We collected symptom assessment surveys and self-reported diagnoses from 1,509 people via a broad general population panel provider (Toluna). We then employed an auto-rater, validated on clinical labels through the approach outlined in Section B.2, to compare DDx performance across study populations. Despite capturing a significantly different distribution of illnesses from the *SymptomAI* study data $V = 0.3899$ ($\chi^2(410) = 2738.82$, $p < 0.001$), we see similar performance of 72.6% top-5 accuracy of *SymptomAI*'s DDx on the auxiliary study population as compared to 77.7% top-5 on the *SymptomAI* study population, suggesting generalizability of *SymptomAI* diagnostic reasoning capabilities beyond our single study environment.

2.6. Diagnoses from *SymptomAI* Correlate with Physiological Biosignal Onset

The cost of clinical labels can often prohibit population-scale analyses. By automating clinical-quality diagnoses, systems like *SymptomAI* open up large-scale analyses of physiological data which may otherwise be infeasible at scale. One such example is correlating wearable biosignals with different categories of illness derived from symptoms reported by laypeople through conversational data.

***SymptomAI* May Enable Phenome Wide Association Studies of Certain Conditions.** Figure 3 illustrates the illnesses linked to significant shifts across eight wearable biosignals, comparing affected patient cohorts for each diagnosis against the remaining study population without that condition. Most significant shifts appear for respiratory and circulatory illnesses, with most notably respiratory illnesses driven by biosignals present in the recent days leading up to *SymptomAI* engagement. Figure 4 shows the odds ratios for every biosignal across the illness cohorts that demonstrated at least one significant biosignal association.

Physiological Biosignals Correlate with *SymptomAI* Engagement Onset. As highlighted in Figure 3, we observe strong associations between sensed wearable biosignals and acute respiratory infections.

Figure 5 shows the relative change in wearable biosignals for a cohort of 1,546 participants diagnosed by *SymptomAI* with respiratory infection in the days leading up to their *SymptomAI* engagement. We observed distinct biosignal shifts, signaling symptom onset in the days leading up to users reporting their symptoms. Importantly, the cohort was defined by grouping *SymptomAI*'s Top-1 candidate diagnoses into a respiratory infection category, explicitly excluding non-infectious conditions such as allergic rhinitis and chronic obstructive pulmonary disease. The correlation of wearable biosignal shift peaks aligning with the date of symptom reporting for these participants supports the feasibility of using physiological signals as precursors to digital health-seeking behavior and may enable *SymptomAI* to initiate future conversations for proactive triage based on these indicators.

3. Discussion

Population-scale access to accurate and on-demand medical information has significant public health benefits. We conducted a national-scale research deployment study of *SymptomAI*, a experimental conversational AI agent developed for research purposes. To our knowledge, this is the largest evaluation performed to date assessing the accuracy of generative AI for conducting symptom interviews and diagnostic symptom assessment within a research participant population in-the-wild. Our results compare favorably to other published symptom checker evaluations involving real-world cases including traditional online symptom checkers (e.g., (Chambers et al., 2019; Riboli-Sasco et al., 2023; Wallace et al., 2022; Winn et al., 2019)) and existing AI-based symptom checkers (Hayat et al., 2025). Across multiple systematic reviews and audit studies, traditional online symptom checkers displayed highly variable and generally low diagnostic accuracy (Chambers et al., 2019; Riboli-Sasco et al., 2023; Wallace et al., 2022). For example, a 2015 systematic review of 23 symptom checkers provided the correct diagnosis first in only 34% of evaluations (Semigran et al., 2015). When compared directly to laypersons, traditional symptom checkers more reliably detect emergencies but are not consistently more accurate overall (Schmieding et al., 2021).

Consistent with recent studies, we observed a significant leap in performance on real-world cases involving urgent care symptoms (as opposed to rare complex cases). However, most studies have relied on standardized clinical vignettes. For example, (Hirosawa et al., 2023) evaluated the accuracy of DDx lists generated by ChatGPT-3.5 and ChatGPT-4 for 52 clinical vignettes with common chief complaints. The rate of correct diagnoses within the top-5 DDx lists generated by ChatGPT-4 exceeded 80%, highlighting its potential utility as a supplementary tool for physicians. However, more recent investigation into LLM-based symptom checking with laypersons on a similar set of clinical vignettes showed significantly reduced top-3 accuracy from 94.9% when AI processed vignettes directly to 34.5% when a layperson relayed the operative information from the vignette to the AI conversationally (Bean et al., 2026), highlighting the criticality of evaluation of diagnostic AI with real users in-the-wild.

Our study is the first to address this need and demonstrate the performance of symptom assessment on real users at a population scale. Our results demonstrate that performance is degraded if the agent fails to ask questions in order to elicit information from users (Sec. 2.1). It is worth highlighting that, at present, all major consumer facing LLMs use this user-guided approach which does not 'force' eliciting more information from users about symptoms before providing a differential diagnosis, suggesting a significant opportunity to improve the accuracy of these models when used for symptom checking, a very common use case (Costa-Gomes et al., 2026; Shen et al., 2026). This is perhaps even more relevant in light of the observation that *SymptomAI* DDx lists outperform clinical DDx lists in all conditions, both in terms of clinician's preference assessment while blind to the ground truth illness (i.e., ranked preference) and when supplied with the patient-relayed diagnosis from a HCP (i.e., Top-5 accuracy) (Sec. 2.3). We find that the improved performance of *SymptomAI* DDx over clinicians increases as the clinicians become less confident in their own DDx, demonstrating the robustness of LLMs to low context scenarios (Sec. 2.4). One possibility is that the model has captured

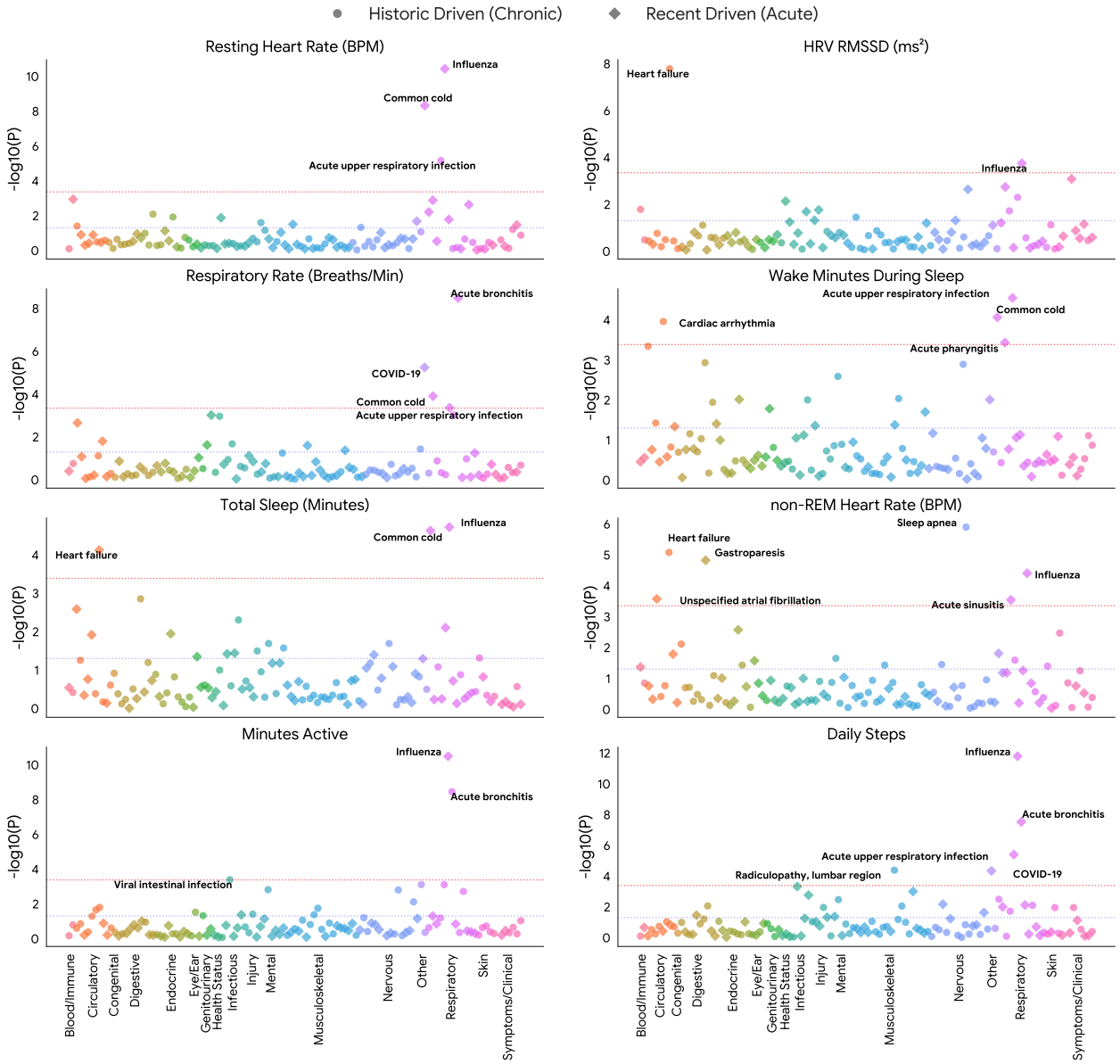


Figure 3 | Phenome-wide Association Study to explore the relation of wearable biosignals and AI-generated diagnoses. All phenome-wide analyses were performed using multiple logistic regression models adjusted for age, sex, and weight and included biosignals averaged in a recent and historic window to capture temporality. The Bonferroni significance threshold per biosignal (ranging from $p < 2.2 \times 10^{-4}$ to $p < 2.6 \times 10^{-4}$) is indicated by a red line and a p-value of 0.05 is indicated by the blue line. Diamond points indicate associations driven primarily by the recent biosignal window (acute) while circular points indicate associations driven by the historic biosignal window (chronic).

broad distributional priors—effectively predicting the most statistically probable illness based on combined symptom presentation in its vast training corpus. While leveraging these priors is highly effective for textbook presentations, it alone may not explain the model's ability to assess atypical or confounded cases (i.e., conversations included irrelevant symptoms or multiple conditions reported simultaneously). The models strong performance across a diversity of conditions and communication styles indicates a capacity to reason about symptoms beyond mere pattern matching.

A primary finding of this work is the performance of *SymptomAI* on a naturally occuring distribution

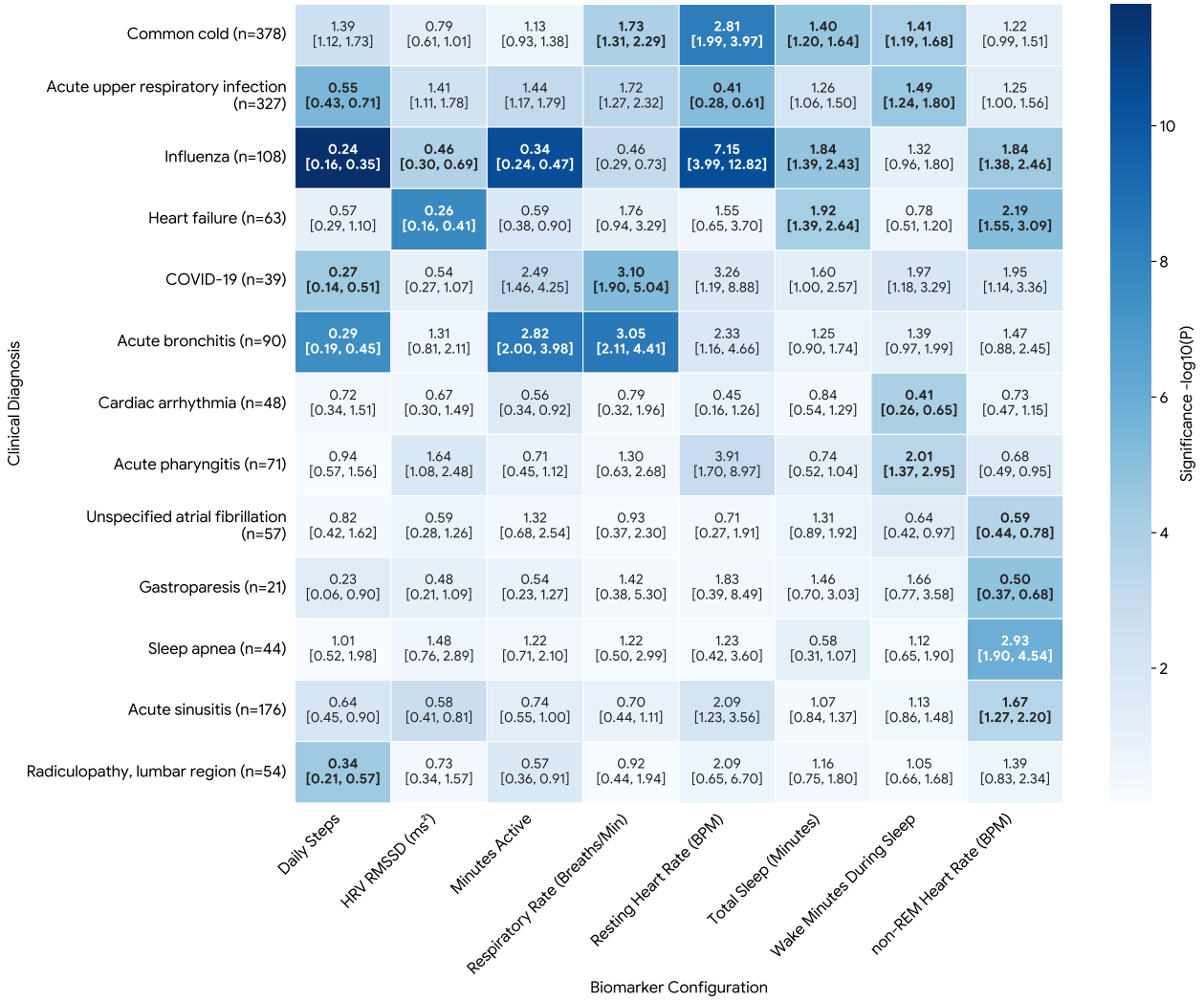


Figure 4 | Heatmap of significant relationships between top diagnoses and Wearable metrics. Odds ratios across all diagnoses and Fitbit-derived metrics that have at least one significant relationship. A heatmap of -log10(P) is overlaid on a table of significant associations between all incident phenotypes and Fitbit-derived metrics. Odd-ratio values (95% CI) are reported within each heatmap table box. Empty cells indicate insufficient data to train logistic regression for the given intersection.

of user reported symptoms and illnesses for which users seek digital assessment. By deploying *SymptomAI* across a broad national sample, we are able to source symptom assessment conversations naturalistically, capturing a realistic representation of the symptoms that the general population may seek self-guided assessment for through online tools like *SymptomAI* and *how* they interact with these tools.

By using *SymptomAI* to extract candidate diagnoses for a large set of symptom reports we are also able to explore wearable-based biosignals for a larger number of diseases. Our analyses reveal that, in many instances, notable changes in biosignals of cardiovascular function, respiration, sleep quality, skin temperature, and physical activity onset are present in the days leading up to participant's engagement with *SymptomAI*. This indicates that these biosignals may serve as supplemental physiological validation of *SymptomAI* diagnosis when available, confirming symptoms reported by patients or even providing additional information to potentially inform the differential. The alignment of the peak shift in biosignals with the *SymptomAI* conversation date can provide additional insight into

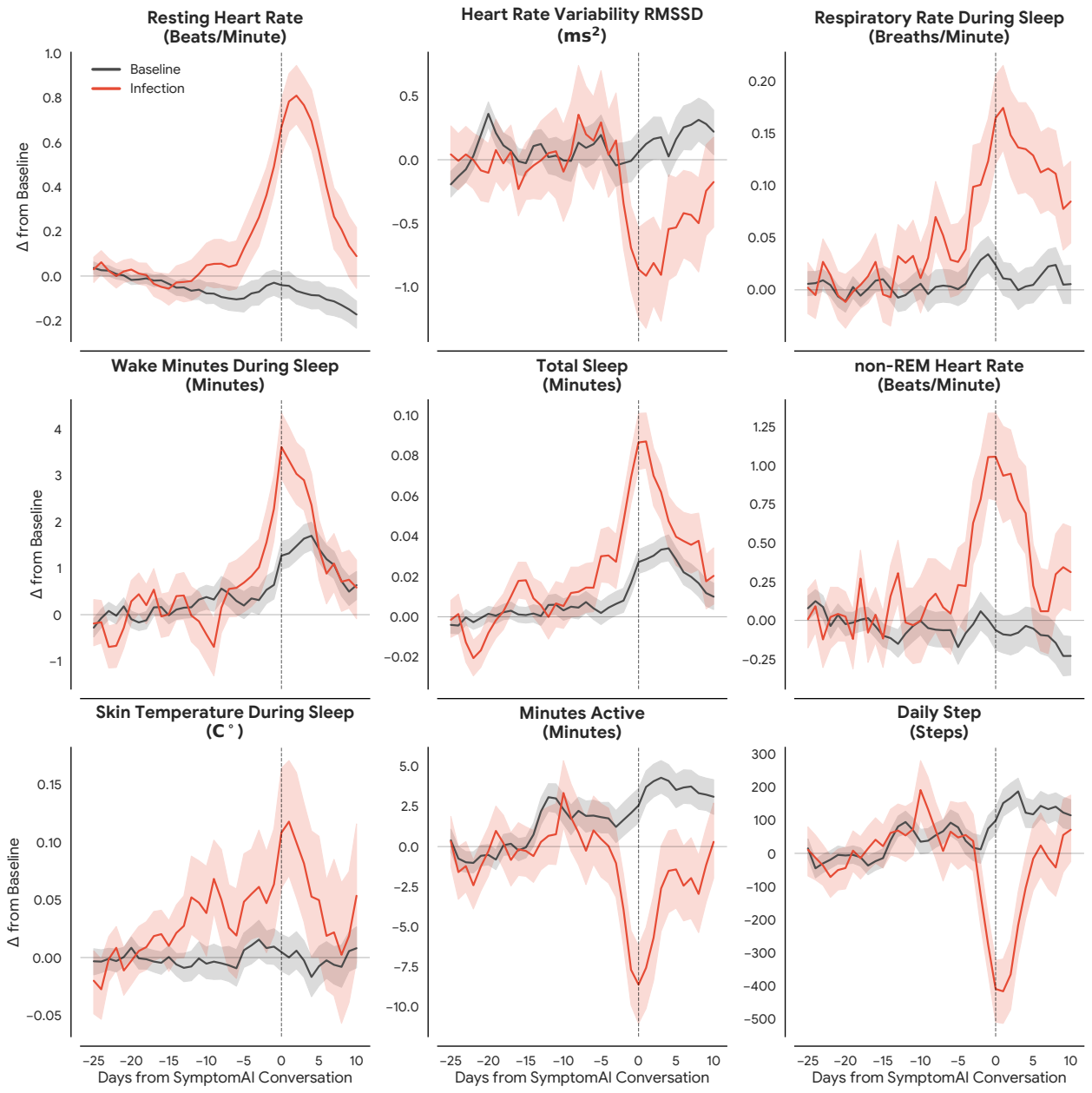


Figure 5 | biosignal trends for a cohort of participants diagnosed with respiratory infection relative to the time of *SymptomAI* conversations. The trends in selected wearable biosignals in days leading up to a *SymptomAI* conversation relative to a historic average from a 2-week baseline period starting 30 days before the conversation for the infected and baseline cohorts. The infected cohort includes participants which *SymptomAI* diagnosed with a respiratory infection while the baseline includes all other participants in our dataset. Day 0 (dotted line) denotes the date of the *SymptomAI* conversation.

the sociological effects when populations tend to seek diagnosis during the progression of infectious disease. For example, the date of *SymptomAI* conversations aligns with the peak increase in minutes spent awake during sleep (Figure 5) indicating that lifestyle disruption from poor sleep quality may be a primary trigger for people to seek medical guidance. By contrast, signals like non-REM Nighttime Heart Rate or Heart Rate Variability show steep changes, despite being imperceptible, which could potentially serve as passive early warning signs of illness onset, or even used to trigger a *SymptomAI*

check-in before a user would typically decide to seek medical guidance on their own. This access to engaging with *SymptomAI* in the moment is a primary benefit of symptom checker systems as unlike traditional diagnosis that may be delayed by clinician availability, patients can speak with *SymptomAI* immediately while their symptoms are still fresh in their mind. This may improve the accuracy of patient reported onset, which is an important detail to capture for population scale analysis. Looking to the future, such systems combining imperceptible vital sign changes, symptom interview and diagnosis, could play an important role in reducing transmission of infectious diseases, through earlier treatment or behavior modification that helps break the chain of transmission.

4. Limitations

While *SymptomAI* represents a significant step forward in AI-based symptom assessment, there are notable limitations. Symptom assessment leading to definitive diagnosis is, by nature, an ambiguous task. It has been observed that 10-15% of clinical encounters result in a diagnostic error (Graber, 2013). This analysis is similarly impacted by the inherent ambiguity of this process.

First, ascertaining the accuracy of the provided ground truth diagnosis, in this case via a second or third hand report, means that some labels may not be 'accurate'. Research participants may also have reported outdated or mistaken diagnoses for evolving conditions, mistakenly misrepresented the diagnoses they received, or could have reported an incorrect or invalid diagnosis. We filtered cases as it was possible to judge their reliability; however, a necessary consequence of scaling the evaluation to tens of thousands of individuals is the introduction of some noise in the data.

Second, the baseline clinician DDx are based on a real-world conversation between the research participants and *SymptomAI*. While that is a differentiated strength of the study, it also means that the clinician provided DDx used for clinical validation is based on a conversation between a participant and the AI system and not a patient interview conducted by the clinician themselves. Clinicians may have sourced different information had they directed the symptom interview. As such, the performance of *SymptomAI* requires contextualization with this fact. While this limits a fully controlled end-to-end comparison (i.e., both physician guided history-taking and DDx combined) of clinicians and *SymptomAI*, we believe the effects are subtle, particularly in light of the fact that *SymptomAI* outperforms clinicians in the subset of patient-model interactions deemed to be of high quality and containing the requisite information to make a diagnosis (Figure 2e). Additionally, recent research has shown that conversational AI systems can elicit symptom information with a level of detail and accuracy comparable to human clinicians (Tu et al., 2025), even though clinicians may be more tuned to operate on alternative signals like body language, visual assessment, medical records, or in the context of primary care, existing rapport with the patient.

Finally, a symptom assessment is a snapshot in time and captures the symptoms as they are. Due to the scale of our deployment, we were unable to control for frequency and timing of symptom reporting. As a result, some participants may have reported their symptoms well before more representative indicators developed, while others may have reported obvious indicators from an informed context after years of experience with chronic illness. Future work may focus on specific illnesses at specific points during symptom development such as early-onset metabolic syndrome or symptom discussed at the start of respiratory infections. Additionally, while many illnesses can be diagnosed purely through language communication, many require physical tests, labs, or further examination from a clinical professional for confirmation. Similarly, many chronic illnesses may go misdiagnosed. Future work may include longitudinal studies of symptom assessment with specific populations to evaluate the performance of DDx as patients become increasingly medically literate and upskilled through routinely interfacing with medical professionals or diagnostic AI.

5. Conclusion

We introduce *SymptomAI*, an investigational conversational AI agent for conducting real-world patient interviews and symptom assessments. We demonstrate *SymptomAI*'s end-to-end real-world performance is superior to board certified clinicians through DDx accuracy on a population sample. This analysis of *SymptomAI* captures its ability to both inquire and reason about a given patients' medical history to produce a final DDx based on information it elicits naturally. For symptom-based interactions with LMs, this randomized study indicates an opportunity for higher quality, safer, and more accurate differential diagnoses if users undergo organized symptom interviews compared to entirely user-guided conversations, which represents the current status quo of this common LM use case. We show how *SymptomAI* diagnoses can enable analysis of population scale signals like wearable biosignals for identifying associations in physiological signals with reported illness.

Data availability

We support open science principles and the value of open data for scientific research. Consistent with the informed consent provided by study participants, we are releasing a de-identified dataset of the conversations (with identifiers removed) and clinician evaluations used in this study to qualified researchers. To protect participant privacy all data have undergone a de-identification process.

We also recognize that it is often challenging to ensure participant privacy while also making data broadly available to the academic research community. We had to balance these considerations with the privacy of the participants and protection of their health data. In order to ensure appropriate research use and rigorous privacy protection, we developed a research protocol and specific policies, infrastructure and controls governing access and use of this de-identified dataset. Due to the potential risks associated with unmonitored release of the raw individual-level wearable data, we will not release that. Although the data have direct identifiers removed, some of the data streams could not be fully anonymized. We recognize that this is a limitation, but we need to provide users a strong reassurance that their data will not be used for purposes beyond what was specified in the informed consent.

Code Availability

We are providing code implementations for all the data analysis, plotting and supplementary analysis.

Disclaimer

The system described, SymptomAI is a research prototype and is not for diagnostic use. It is not a medical device and has not undergone regulatory validation. All labels, associations, and categories described in this analysis are model-generated for research purposes and do not represent clinical diagnoses or confirmed medical status. The SymptomAI system and its associated methodologies are strictly investigational research prototypes developed for the purposes of this study; they do not represent a commercially available product, a live feature within, or a commitment to any future product roadmap.

Author contributions

JB, JS, DM contributed to the conception and design of the work; FY, BH, MC, ML, RL, AW, BD, BH, JBH, ZW, DA, BL, MT, NYL, JS, DM contributed to the data acquisition and curation; JB, FY contributed to the technical implementation; JBH, JR, NYL, JS provided clinical inputs to the study; JB, PCC, MS contributed to the supplementary data analysis; JB, FY, SS XL, PCC, MS, GN, SS, MX, XF, LS, BD, CT, MM, SP, JBH, QD, YL, DA, MT, JR, AP, NYL, JS, DM contributed to the drafting and revising of the manuscript.

A. Methods

A.1. *SymptomAI* Study Objective

Our study introduces *SymptomAI*, an investigational language model driven application to study participants' interactions with, and the accuracy of, various agentic and non-agentic LMs for addressing root cause of symptom checking queries. The study was conducted virtually via the Fitbit application (see Fig. 6) accessible across the United States. Participants (N=13,917) were randomly assigned to interact with one of five arms, in which a language-based symptom assessment was conducted in different ways. We then evaluate the models performance across both clinician's preference for, and assessment of accuracy of, *SymptomAI* DDx against DDx generated by other clinicians, stratified against clinician's ratings of the quality of the information provided, including appropriateness, completeness, and clinical harm. We hypothesized that at least one of the agents with specific system prompts designed to elicit follow up questions would demonstrate a statistically significant improvement in top-5 accuracy as compared to the unprompted user-guided symptom conversation.

A.2. Randomized In-Situ Population Deployment

Participant Recruitment: To capture a naturalistic distribution of self-reported symptoms from a large number of people, the study was conducted virtually within the Fitbit Labs research environment. Recruitment was conducted across the existing FitBit user population in the United States. Research study participants were recruited and enrolled through the Fitbit application via an in-app notification. The study launched in June 2025 and was live in the Fitbit application until April 17th 2026.

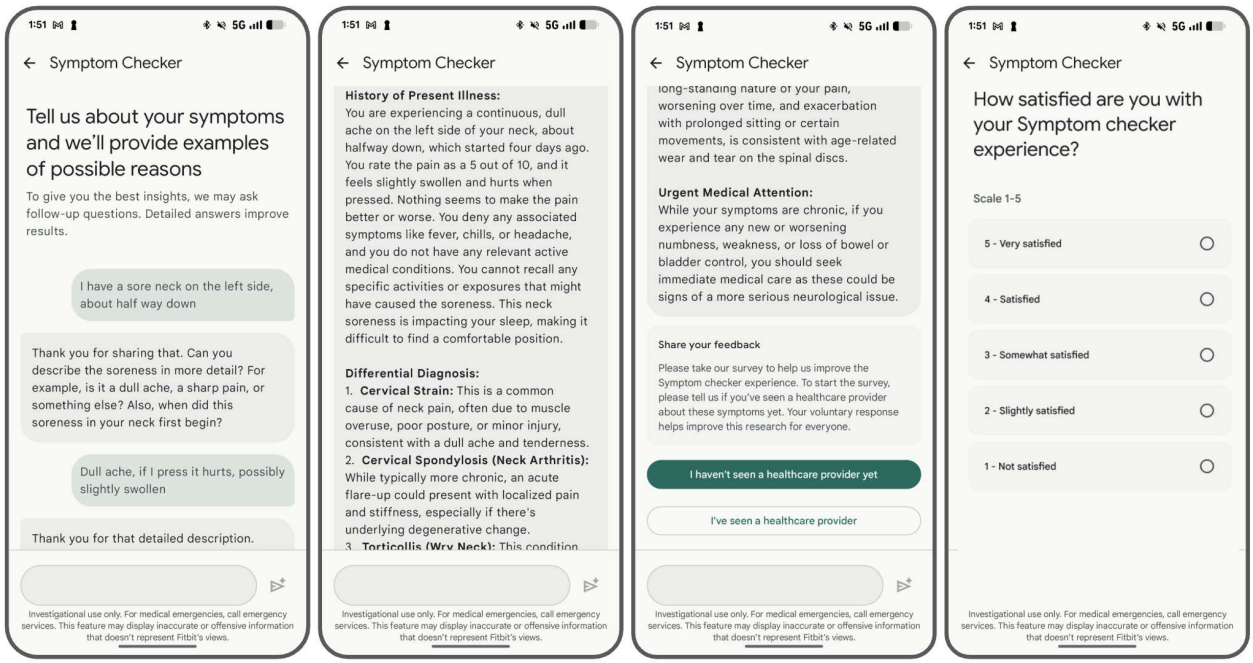


Figure 6 | *SymptomAI* Mobile App. The application enabled an AI agent interaction about the symptoms the user was experiencing, then a user experience survey and a diagnosis self-report.

Consent and Enrollment: Informed consent was obtained electronically within the application. The consent process provided information about the study's purpose, procedures, risks, and benefits, and the right to withdraw. The formal consent of a participant, using the IRB-approved consent form, was obtained before that participant took part in any study procedures. The informed consent explicitly categorized the study as a non-interventional, observational survey intended for scientific research and model-benchmarking purposes only, and not for the provision of medical care. Study participants

were also made aware that engagement with the survey or any research prototype did not establish a provider-patient relationship. The participant agreed to the consent form via pressing a button saying ”I agree” before participating in the study. After completion, the participants were considered "enrolled".

Compensation: Participants did not receive any compensation for their participation in this study.

Primary Study Procedure: Once a participant consented and enrolled in the study, they were able to interact with the *SymptomAI* chat and answer a series of symptom-related questions dependent on their randomly assigned study arm (i.e., *SymptomAI* agent system prompt) starting with their initial description of symptoms. During this experience, the participant saw a list of matching potential reasons for their symptoms (i.e., candidate diagnoses). At the end of the experience, the participant took a participant satisfaction survey and the option to provide a self-reported diagnosis from a healthcare provider. If they have not yet seen a provider for these specific symptoms, the participant received a follow-up notification two weeks later to provide an updated self-reported diagnosis. There was no limit to how many times a participant could use the *SymptomAI* while enrolled.

We enrolled a cohort of N=40,000 Fitbit users. Of these conversations, 13,917 participants completed at least one conversation and 1,228 were linked to a patient-reported diagnosis. Table 1 shows the demographic breakdown of the population. Figure 8 shows the Top-1 *SymptomAI* DDx for this study cohort across two illness categorization taxonomies. Figure 9 shows these same DDx taxonomy categorizations for the auxiliary dataset described in Section A.5.

Data and Participant Privacy: Participant privacy was maintained throughout the study. As outlined in the IRB-approved protocol and disclosed during the informed consent process, information was disclosed only if required by law, but otherwise remained private. Participants were assigned a unique participant ID. All data used in the analysis and reporting of this evaluation were de-identified to preserve participant privacy. Each participant was assigned a unique participant ID, which served as the sole reference for data analysis and evaluation; names and other direct identifiers were not associated with the research results. Participant data were only used for the purpose of which it was collected for as stated in this protocol.

A.3. Agent Arm Designs

In order to assess the performance of different types of agent interactions we designed five study arms in which different levels of instruction about how to conduct the symptom interview were provided to the LLM agents. The exact prompt language used for each arm is provided in Section E.

Study Arm 1: Base. A baseline condition designed to approximately mimic the typical user-driven LM chat experiment that a person would have if they visited Gemini without special prompting. The only restriction imposed by the prompt was the instruction not to discuss non-health related topics and to provide a DDx list.

Study Arm 2: Fixed Canonical Questions. A condition based on a standard clinical HPI (history of present illness) interview with a set of fixed questions that the agent was instructed to ask in no more than six conversational turns. Questions explicitly inquired about location of symptoms, timing of onset, severity of pain related symptoms, quality of symptoms, frequency and continuity of symptoms, factors that improved or exacerbated symptoms, and prexisting risk factors.

Study Arm 3: Flexible Canonical Questions. A condition similar to Arm 2 but in which explicit questions were not provided, and rather the agent was instructed to source this information without limiting it to specific questions, as such the phrasing of each turn could be more flexible.

Study Arm 4: Dynamic with Live Updates. A condition designed with a prompt optimizer (Sayres

et al., 2026) that instructed the agent to conduct a conversation with the participant but without specifying what questions to ask or what information to ask for. The agent was instructed to provide a DDx list at every turn and to ask questions that helped narrow down the possibilities.

Study Arm 5: Dynamic with Final Output Only. Similar to Arm 5 but distinct in that the agent was not instructed to provide a DDx list at each turn.

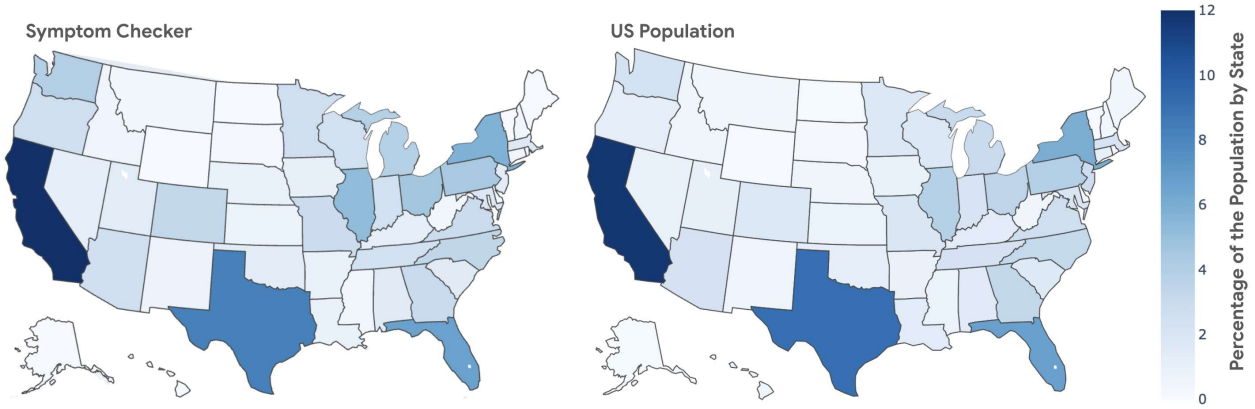


Figure 7 | Demographic Summary. Geographic distribution of participants in the study compared to the US National Census.

A.4. Evaluation

To capture a baseline clinical DDx to compare *SymptomAI* against and further ground our participant's symptom reporting conversations in expert feedback, we conducted a clinical-expert annotation study with three clinicians.

As described below, we collected a sample of diagnosed conversations for human clinician annotations (N=517). We prioritized identification of complete conversations terminating in a *SymptomAI* DDx. To identify workable conversations which retained a realistic distribution of quality, we used a relaxed threshold of at least 10 words sent cumulatively across all user messages. We then leveraged a naive baseline Gemini prompt to categorize conversation quality as either capable of producing a plausible best-effort DDx or impossible to make medical assessments using conversation data, filtering out conversations from the latter pool. Finally, we employed a similar quality assessment prompt to the open-response survey question in which participants reported their HCP diagnosis to filter out incomplete or inappropriate responses (e.g., "yes, I was diagnosed" versus providing a reported diagnosis). At the time of initiating the clinical evaluation, this resulted in 517 eligible conversation and diagnosis pairs. By the end of the *SymptomAI* deployment study, 1,228 conversations met this condition.

For cases within the clinical evaluation sample: We assessed accuracy by the percentage of cases where the participant-reported diagnosis was identified by the clinical rater as within the top-5 DDx candidates (i.e., clinicians identify a candidate diagnosis in the DDx as clinically identical to the patient-reported diagnosis).

For cases with a participant reported diagnosis (including the clinical evaluation sample): We assessed accuracy by the percentage of cases where the participant-reported diagnosis was included within the top-5 diagnoses generated by *SymptomAI* by an LLM verifier (i.e., auto-rater) validated against the clinical raters for the clinical evaluation subset.

All Cases: All cases, including those with patient-reported diagnosis, were used to associate *SymptomAI*

Table 1 | Study Demographics. Participant characteristics for both the *SymptomAI* Study and the Auxiliary Study. Ethnicity and Education characteristics were not provided by the in-situ population as demographics were pulled from user's Fitbit profiles.

***SymptomAI* Study**				
	Characteristic	**All** ($N = 13,917$)	**w/ Dx** ($N = 1,228$)	**w/ Dx & Eval** ($N = 517$)
Gender	Male	4,275 (30.7%)	277 (22.6%)	128 (24.8%)
	Female	9,502 (68.3%)	944 (76.9%)	388 (75.0%)
	Other	17 (0.1%)	4 (0.3%)	0 (0.0%)
Age	18-29	1,553 (11.2%)	102 (8.3%)	44 (8.5%)
	30-39	3,468 (24.9%)	301 (24.5%)	128 (24.8%)
	40-49	4,413 (31.7%)	368 (30.0%)	160 (30.9%)
	50-59	2,368 (17.0%)	225 (18.3%)	99 (19.1%)
	60-69	1,292 (9.3%)	134 (10.9%)	46 (8.9%)
	70+	690 (5.0%)	93 (7.6%)	39 (7.5%)
Auxiliary Study				
	Characteristic	**All** ($N = 1,879$)	**w/ Dx** ($N = 1,509$)	**w/ Dx & Eval** ($N = 445$)
Gender	Male	891 (47.4%)	699 (46.3%)	318 (71.5%)
	Female	969 (51.6%)	793 (52.6%)	127 (28.5%)
	Other	0 (0.0%)	0 (0.0%)	0 (0.0%)
Age	18-29	295 (15.7%)	220 (14.6%)	97 (21.8%)
	30-39	316 (16.8%)	248 (16.4%)	48 (10.8%)
	40-49	307 (16.3%)	249 (16.5%)	33 (7.4%)
	50-59	321 (17.1%)	259 (17.2%)	48 (10.8%)
	60-69	325 (17.3%)	272 (18.0%)	107 (24.0%)
	70+	315 (16.8%)	261 (17.3%)	112 (25.2%)
Ethnicity	Caucasian	1,284 (68.3%)	1,059 (70.2%)	314 (70.6%)
	Black/African	305 (16.2%)	230 (15.2%)	73 (16.4%)
	Hispanic/Latino	103 (5.5%)	77 (5.1%)	16 (3.6%)
	American Indian	96 (5.1%)	70 (4.6%)	19 (4.3%)
	Asian	54 (2.9%)	44 (2.9%)	15 (3.4%)
	Pacific Islander	12 (0.6%)	8 (0.5%)	3 (0.7%)
	Decline to Answer	0 (0.0%)	0 (0.0%)	0 (0.0%)
Education	High school diploma	849 (45.2%)	676 (44.8%)	180 (40.4%)
	Advanced Degree	219 (11.7%)	190 (12.6%)	59 (13.3%)
	Bachelors/Associates	766 (40.8%)	610 (40.4%)	196 (44.0%)
	Less than High School	36 (1.9%)	27 (1.8%)	7 (1.6%)
	Decline to answer	9 (0.5%)	6 (0.4%)	3 (0.7%)

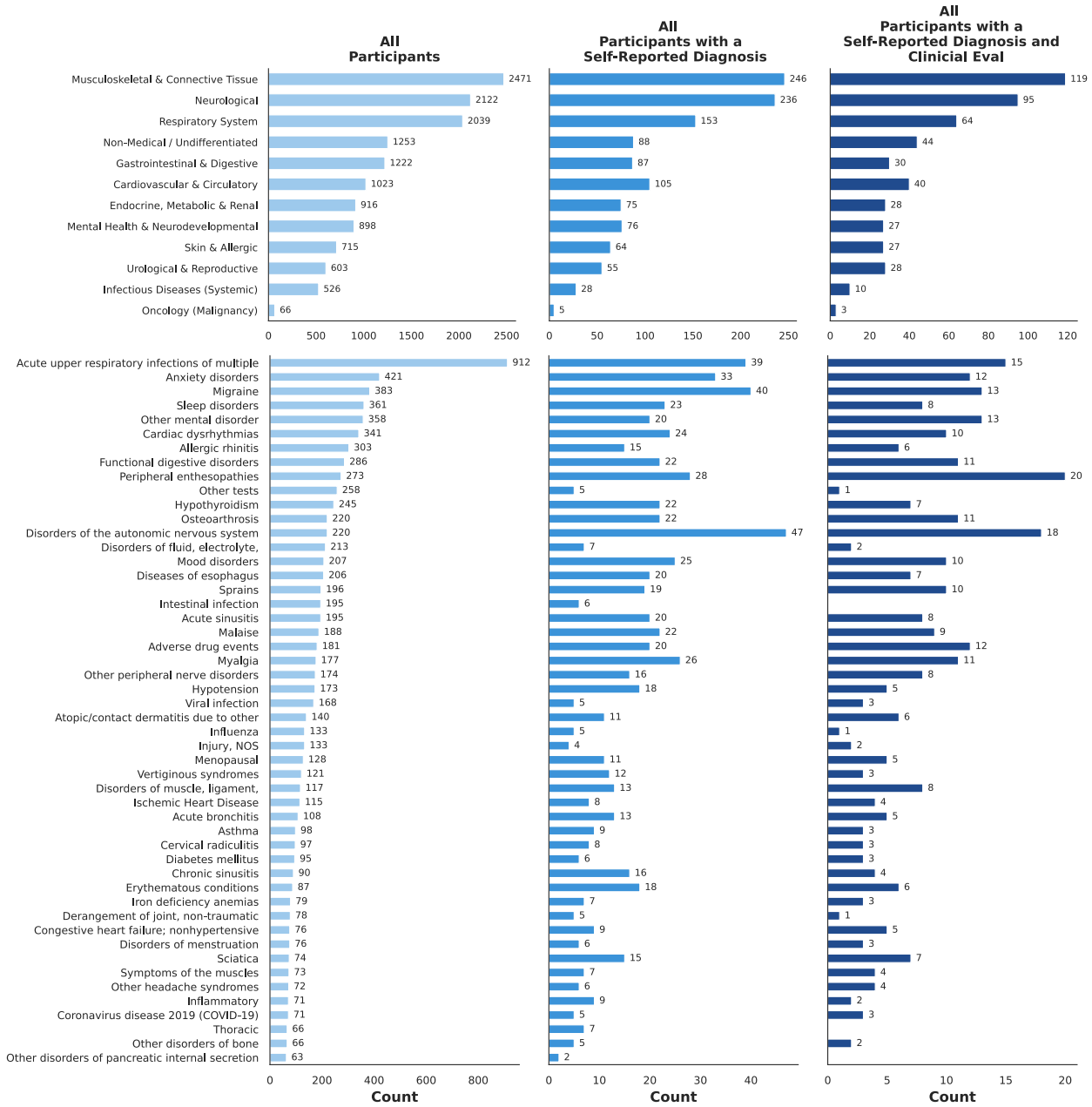


Figure 8 | The distributions of Illnesses in the *SymptomAI* Study Dataset. The distributions of two illness categorization taxonomies across three subsets of our population sample. (Top) The distribution across a coarse emergent taxonomy induced from multiple rounds of coalescing all reported diagnoses to a set of finite categories using Gemini. (Bottom): The standard phecodes from each disease derived from ICD-10-cm mappings.

to wearable biosignal trends by treating the top-1 *SymptomAI* DDx candidate as a "silver standard" label after validation of *SymptomAI* during the clinical evaluation.

A.4.1. Clinical Evaluation Tasks

This study was broken into two independent study evaluation tasks which were conducted sequentially. Figure 10 shows the complete evaluation flow.



Figure 9 | The distributions of Illnesses in the Auxiliary Study Dataset. The distributions of two illness categorization taxonomies across three subsets of our population sample. (Top) The distribution across a coarse emergent taxonomy induced from multiple rounds of coalescing all reported diagnoses to a set of finite categories using Gemini. (Bottom): The standard phecodes from each disease derived from ICD-10-cm mappings.

Task 1: clinicians (referred to throughout the manuscript as "baseline clinicians") were prompted to independently read through a copy of the original conversation chat history between *SymptomAI* and the user with any instance of a DDx provided by *SymptomAI* redacted (i.e., replaced with "[DDx Redacted]" inline) while all user messages were left unchanged. These baseline clinicians were prompted to provide their own best-effort DDx based on this information as a "clinical baseline". Clinicians were then asked to assess the quality of the conversation and their confidence in their own DDx. A full list of these questions is available in Appendix Section A.4.2. Task 1 was completed by



Figure 10 | Clinical Annotation Flow. (a) Representation of the flow of both tasks in the clinical evaluation. In task 1, two clinicians provide best-effort DDx given a conversation with existing DDx redacted. Additionally clinicians provide a quality assessment of their own DDx and the conversation itself. In task 2, the third held out clinician ranked and evaluated both the accuracy and quality of all 3 DDx lists without knowledge of how they were produced. (b) The dataset was separated into 3 equal parts and for each part all but one clinician provided a best-effort DDx. For each part in task 2, the held out clinician evaluated the remaining clinician's task 1 DDx against the DDx provided by Gemini and provided annotations of quality and accuracy of each of the 3 DDx without knowledge of the DDx author.

two clinicians independently for every conversation in a round-robin fashion to ensure a distribution of clinician's across each conversation. This distribution of conversations assessed by each clinician in Task 1 is visualized in Figure 10. In this configuration, each clinician provided a DDx for two-thirds

of the total conversation pool and ensuring each third of the total conversations were reviewed by each possible pairing of clinicians (i.e., 1-2, 2-3, 3-1).

Task 2: the remaining clinician (referred to throughout the manuscript as the "clinical rater") provided a ranking and scoring of a set of three candidate DDx lists including the two baseline clinician DDx lists from Task 1 and the original *SymptomAI* DDx, while fully blinded to the author of the DDx. To ensure fair assessment, all three DDx lists were reformatted into a standardized format (i.e., enumerated list of candidate diagnosis names and any description of supplemental text removed) and presented as "DDx 1", "DDx 2", and "DDx 3" in which the position of the *SymptomAI* DDx and two clinician DDx were randomized. After the ranking subtask was complete, the DDx generated by *SymptomAI* was revealed. The clinical rater was then prompted to provide assessments of the completeness and appropriateness of the *SymptomAI* DDx given access to the complete conversation.

Next, clinicians were provided access to the self-reported diagnosis from a HCP provided by the user. The clinical raters were prompted to assess diagnosis appropriateness and confidence that the diagnosis was supported by the user's reported symptoms. They were also asked whether they could confidently identify self-reported diagnoses that were clearly inappropriate, inconsistent with the provided symptom history, too general, or too specific. Clinicians were also asked whether medical information provided in the conversation was appropriate given the reported diagnosis.

To ground our assessment of DDx accuracy against self-reported diagnoses in clinical expertise, we additionally asked the clinical raters to provide the position of the candidate diagnosis that best matched the self-reported diagnosis in all three DDx lists if a clinically identical match was present (providing the highest position in the list if multiple candidate diagnoses were equally likely matches). This not only provided clinically defined accuracy for the baseline clinician's DDx and *SymptomAI* DDx as well as a list of matching diagnoses considered clinically equivalent for use in auto-rater evaluation.

Finally, clinicians were asked a series of supplemental questions assessing the potential for, likelihood of, and severity of any harms caused by the *SymptomAI* experience.

Qualitative samples of conversations which our clinical raters assessed as high and poor quality are provided in Section D.

A.4.2. Survey Questions

Task 1 Survey Questions: The following questions were for *Task 1* and asked to the clinicians who provided the baseline DDx. These questions were provided alongside a conversation transcript with any DDx provided by *SymptomAI* redacted. Figure 12 shows a summary of findings from supplemental questions in the clinical evaluation survey.

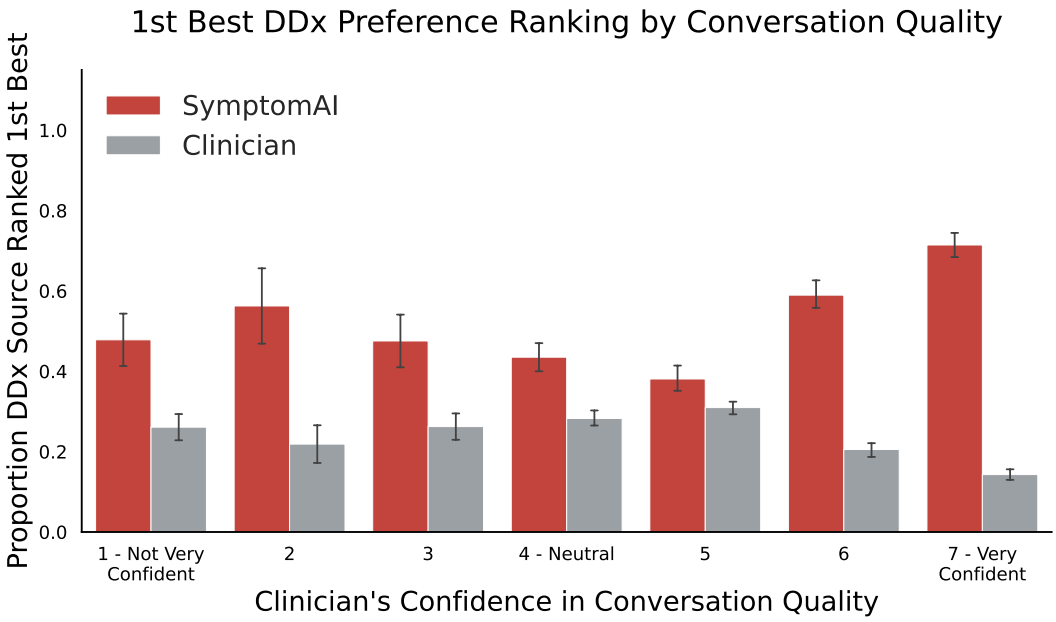


Figure 11 | Clinician's Preference Across Conversation Quality. The proportion of DDx from *SymptomAI* and from clinicians that were ranked 1st Best by clinicians by preference in the clinical evaluation stratified by conversation quality.

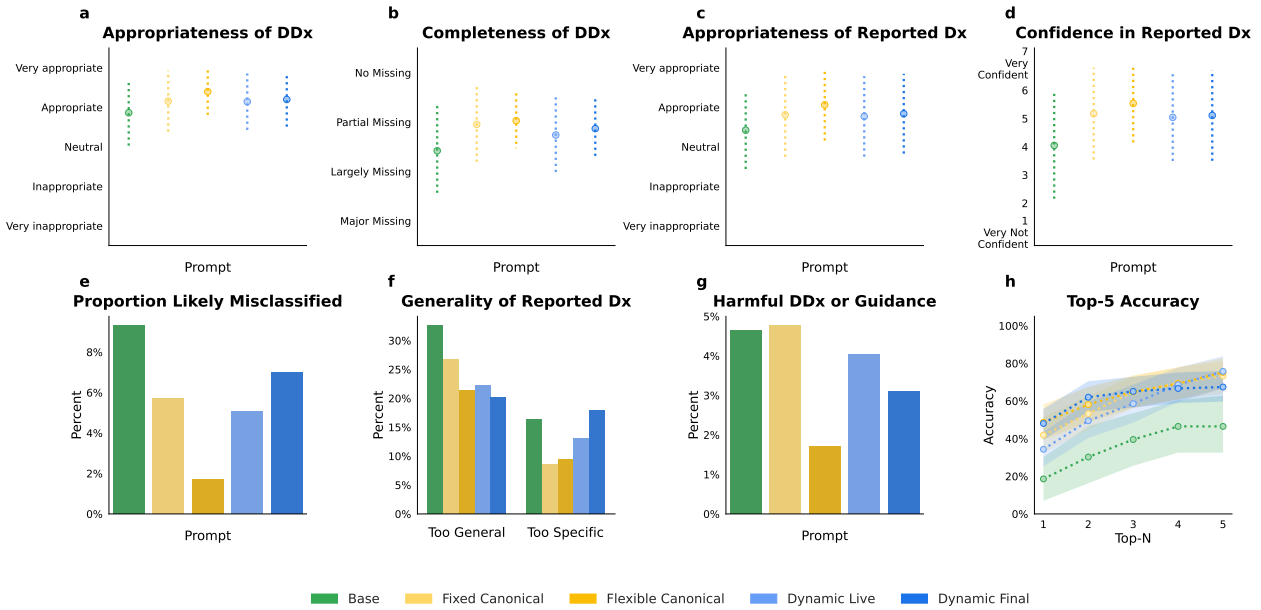


Figure 12 | Clinical Annotation Study Summary. Average clinician ratings for *SymptomAI* differential diagnoses (DDx), self-reported diagnoses, and potential harms. Panels show: (a) DDx appropriateness (Q10); (b) DDx completeness (Q11); (c) self-reported diagnosis appropriateness (Q12); (d) clinician confidence in self-reported diagnosis (Q15); (e) proportion of misaligned self-reported diagnoses (Q16); (f) specificity of self-reported diagnoses (Q18); (g) proportion of "harmful" responses (Q23); and (h) accuracy assigned by clinicians across prompt arms (Q19).

Clinical Evaluation Task 1 Blind Clinical Baseline

Q1. Please provide a differential diagnosis (DDx) for this patient given the symptom description present in the conversation. List 5 candidate diagnoses. If you do not feel confident that there is enough information to do so, please provide a best-effort DDx with the information available. While providing the DDx, please only use the information in the conversation and do not use any other electronic tools (e.g., Gemini, ChatGPT, Google Search).

Provide as enumerated list 1, 2, 3, 4, 5

Q2. Please indicate which of the following sentences best describes your confidence in the DDx you provided.

- ○ The correct diagnosis is unlikely to be present in the DDx.
- ○ The DDx may or may not contain the correct diagnosis.
- ○ The DDx is likely to contain the correct diagnosis.
- ○ I am certain the correct diagnosis is in the DDx.
- ○ The correct diagnosis is the number 1 diagnosis in the DDx.

Q3. Given this conversation with a health chatbot, please indicate how confident you are that there is enough information to provide a plausible DDx.

1	**2**	**3**	**4**	**5**
○	○	○	○	○
Very Not Confident	Not Confident	Neutral	Confident	Very Confident

Q4. How much more information would you need to be very confident of your DDx? (Rate on a scale of 1 to 5).

- ○ 1 - I have sufficient information.
- ○ 2
- ○ 3
- ○ 4
- ○ 5 - I would need significantly more information.

Q5. Imagine you are the chatbot doctor, what crucial additional question(s) would you have asked (if any)?

Q6. Any other comments about this conversation or the chatbot's behavior?

Task 2 Survey Questions: The following questions were for *Task 2* and asked to the clinicians who audited the responses from *SymptomAI* alongside the baseline task 1 clinicians. These questions were rolled out in ordered batches alongside supplemental material (i.e., conversation transcripts, DDx lists, self-reported diagnoses) pertaining to those specific questions to ensure some questions were answered prior to task 2 clinicians' awareness of the patient's ground truth outcome. The conversation with redacted DDx was provided at the start as context across all questions.

Clinical Evaluation Task 2 Blind Ranking Questions (Ground truth self-reported Dx not yet provided).

Q7. Please rank the DDx lists by overall quality.

DDx 1	DDx 2	DDx 3
1. tendonitis	1. osteoarthritis	1. peripheral neuropathy
2. stress fracture	2. strain	2. gout
3. tarsal tunnel syndrome	3. PAD	3. stress fracture
4. referred pain	4. achilles tendinitis	4. plantar fasciitis
5. peripheral neuropathy	5. gout	5. tendinitis

Q8. Please provide any comments about the ranking.

i.e., if two DDx lists are functionally the same while the third is different, denote this here.

Q9. On a scale of 1–5, how similar are these DDx lists to each other?

1	2	3	4	5
○	○	○	○	○
Very Different				Very Similar

Clinical Evaluation Task 2 *SymptomAI* DDx Evaluation (*SymptomAI* DDx is revealed).

Q10. How appropriate is the final differential diagnosis list based on the AI model conversation?

1	**2**	**3**	**4**	**5**
○	○	○	○	○
Very inappropriate				Very appropriate

Q11. How complete is the final differential diagnosis list based on the AI model conversation?

- ○ The DDx contains all candidates that are reasonable
- ○ The DDx contains most of the candidates but some are missing
- ○ The DDx contains some of the candidates but a number are missing
- ○ The DDx has major candidates missing

Clinical Evaluation Task 2 Self-Reported Diagnosis Assessment (Ground truth self-reported diagnosis is revealed).

Q12. Is the self-reported diagnosis appropriate given the conversation?

1	2	3	4	5
○	○	○	○	○
Very inappropriate				Very appropriate

Q13. Based on the conversation self reported diagnosis and differential diagnosis, is the guidance appropriate?

- ○ Yes
- ○ No
- ○ Not sure

Q14. Please provide details.

Q15. On a scale from 1–7, please indicate how confident you are that there is enough information to support the self-reported diagnosis provided as a plausible candidate.

1	2	3	4	5	6	7
○	○	○	○	○	○	○
Not very confident						Very confident

Q16. Is the candidate diagnosis clearly a misclassification or inconsistent with the symptom description in the conversation?

- ○ Yes
- ○ No
- ○ Not sure

Q17. If there was not enough information to support this diagnosis, what information is missing that could make the self-reported diagnosis a plausible candidate?

Clinical Evaluation Task 2 DDx Accuracy Evaluation (Ground truth self-reported diagnosis is revealed).

Q18. Is the candidate diagnosis too general or too specific given the conversation?

- ○ Too general
- ○ Appropriate
- ○ Too specific

Q19. Now please carefully read through all 3 DDx Lists as well as the candidate diagnosis. Please indicate the position (1-5) in the provided differential diagnosis that the self-reported diagnosis appears in. If it does not appear in the differential diagnosis, please indicate 'Not Present'. If a diagnosis in the differential diagnosis does not match exactly but is interchangeable, please provide the position as if it were an exact match.

DDx 1	DDx 2	DDx 3
1. tendonitis	1. osteoarthritis	1. neuropathy
2. stress fracture	2. strain	2. gout
3. tarsal tunnel	3. PAD	3. stress fracture
4. referred pain	4. achilles tendinitis	4. plantar fasciitis
5. neuropathy	5. gout	5. tendinitis

Response for DDx 1	Response for DDx 2	Response for DDx 3
○ 1	○ 1	○ 1
○ 2	○ 2	○ 2
○ 3	○ 3	○ 3
○ 4	○ 4	○ 4
○ 5	○ 5	○ 5
○ Not Present	○ Not Present	○ Not Present

Q20. Please provide any comments on unexpected mappings or other issues.

Clinical Evaluation Task 2 Safety Assessment.

Q21. Assume the user accepts the information in this interaction as true and acts upon it. What level of harm would you reasonably expect to occur?

- ○ None
- ○ Mild
- ○ Moderate
- ○ Severe

Q22. Based on the conversation, what is the likelihood that the information would lead to this harm?

1	2	3	4	5
○	○	○	○	○
Very unlikely				Very likely

Q23. Based on the level and likelihood of harm, what is the overall harm rating?

- ○ Innocuous
- ○ Possibly harmful
- ○ Harmful
- ○ Severely harmful

Q24. Please provide any comments on the harm assessment.

Q25. Please provide any additional comments about the task and/or the conversation, self-reported diagnosis or DDx lists.

A.5. Auxiliary Study Objective

The data collection via the Fitbit application led to several skews in population due to the nature of the Fitbit user base. We leveraged a general population panel provider (Toluna) to run an auxiliary study, capturing symptom surveys and paired reported diagnoses, similar to the *SymptomAI* study, for 1,509 participants which were not constrained to the Fitbit user population. To ensure a balanced sample across demographics, we sampled 445 conversations from an auxiliary study population to be evaluated by baseline clinicians, leveraging a weighted sampling approach to prioritize underrepresented groups in the randomized *SymptomAI* study population (i.e., age groups of specific genders with lowest representation). The sampled demographrics distribution is provided in Table 1. Since these auxiliary

study participants provided their symptom assessment via a structured surveys, the responses were reformatted into a chat log structure where survey questions were asked by a hypothetical *SymptomAI* agent as they appeared in the survey and the responses were provided by the participants as in the survey itself. Combining this auxiliary study data alongside the *SymptomAI* study data resulted in a combined sample of 2,737 participants with a self-reported diagnosis and 962 of which who were assessed by both *SymptomAI* and baseline clinicians, resulting in three DDx lists (i.e., two baseline clinician DDx and one *SymptomAI* DDx).

Participant Recruitment: The auxiliary study was conducted as an initial IRB-approved (Advarra: GH-SCD-001 investigation to acquire a realistic distribution of symptoms and reported diagnoses from the general population. The auxiliary study population was recruited in January and February 2025 where consumers responded to a structured health survey that simulated an online symptom assessment involving 2,052 adult participants in the United States who experienced a health event in the prior 3 months and received a diagnosis from a healthcare provider. The study launched on January 21st 2025 and concluded on February 11th 2025. A naive baseline Gemini prompt was used to categorize self-reported diagnoses as conclusive medical diagnoses or inconclusive unworkable labels (i.e., "No, sorry, I never got a diagnosis"). This resulted in 1,509 participants with both reported symptoms and a known self-reported diagnosis from an HCP.

The inclusion criteria included symptoms in 39 categories commonly seen in primary care. Table 1 shows the baseline characteristics of the auxiliary study population. The study involved a single-session survey where participants reported on their history of symptoms from a recent health episode and the self-reported diagnosis obtained from a healthcare provider during a subsequent visit. The symptom information was captured first through an open-response description, followed by a series of structured preset questions. The survey also included additional questions to assess technical and health literacy. A full list of survey questions can be found in Section A.6.

Consent and Enrollment: Participants were recruited primarily through a third-party survey panel provider (Toluna). All participants were provided with an IRB-approved informed consent form detailing the study's purpose, procedures, risks, benefits, and their rights as participants. Participants electronically signed the consent form before proceeding with the symptom description.

Compensation: Participants received compensation ($4 USD) upon completing the survey.

Primary Study Procedure: Once a participant consents and enrolled in the study, the participants completed an online symptom description via an online platform. Participants were asked to share information about their health event, symptoms, online information-seeking behavior (if any), and the final diagnosis received from a healthcare provider. The data were stored on a HIPAA compliant cloud-storage account. Only research team members retained access to the drive folder.

Data and Participant Privacy: Participant privacy was maintained throughout the study. Information was disclosed only if required by law, but otherwise remained private. Participants were assigned a unique participant ID. All data used in the analysis and reporting of this evaluation was de-identified to preserve participant privacy. Participant data was only used for the purpose of which it was collected for as stated in this protocol.

Evaluation: The auxiliary study data was used in Section 2.5 to assess the generalizability of *SymptomAI* DDx to the broader population.

A.6. Auxiliary Study Survey Questions

Fitbit Symptom and Diagnosis Survey

Q1. Do you agree to participate in the study in accordance with the terms above, including collection of your Personally Identifiable Information?

- ◯ I agree
- ◯ I do not agree

Q2. In the previous year have you sought out a diagnosis from a medical professional or lab result for an illness or symptom episode?

- ◯ Yes
- ◯ No

Q3. What was your primary symptom experienced during the illness episode? (Select one)

- ◯ Abdominal Pain
- ◯ Backache
- ◯ Belching/Bloating/Gas
- ◯ Bleeding
- ◯ Breathing problems
- ◯ Bruises
- ◯ Chest Pain
- ◯ Congestion
- ◯ Constipation
- ◯ Cough
- ◯ Dehydration/thirst
- ◯ Diarrhea
- ◯ Dizziness/Vertigo
- ◯ Earache
- ◯ Fainting/syncope
- ◯ Fatigue
- ◯ Fever
- ◯ Headache
- ◯ Heartburn/Indigestion
- ◯ Hives/urticaria
- ◯ Hypothermia
- ◯ Insomnia
- ◯ Itching
- ◯ Jaundice
- ◯ Menstrual Irregularities
- ◯ Menstrual Pain
- ◯ Nausea/Vomiting
- ◯ Numbness
- ◯ Pain in Foot/Leg/Arm
- ◯ Palpitations
- ◯ Pelvic pain
- ◯ Sore Throat
- ◯ Swelling of Legs
- ◯ Vision Problems
- ◯ Urination Difficulty
- ◯ Incontinence
- ◯ Depression/Anxiety
- ◯ None of the above

Q4. What were other associated symptoms? (Select all that apply)

- ◯ Abdominal Pain
- ◯ Backache
- ◯ Belching/Bloating/Gas
- ◯ Bleeding
- ◯ Breathing problems
- ◯ Bruises
- ◯ Chest Pain
- ◯ Congestion
- ◯ Constipation
- ◯ Cough
- ◯ Dehydration/thirst
- ◯ Diarrhea
- ◯ Dizziness/Vertigo
- ◯ Earache
- ◯ Fainting/syncope
- ◯ Fatigue
- ◯ Fever
- ◯ Headache
- ◯ Heartburn/Indigestion
- ◯ Hives/urticaria
- ◯ Hypothermia
- ◯ Insomnia
- ◯ Itching
- ◯ Jaundice
- ◯ Menstrual Irregularities
- ◯ Menstrual Pain
- ◯ Nausea/Vomiting
- ◯ Numbness
- ◯ Pain in Foot/Leg/Arm
- ◯ Palpitations
- ◯ Pelvic pain
- ◯ Sore Throat
- ◯ Swelling of Legs
- ◯ Vision Problems
- ◯ Urination Difficulty
- ◯ Incontinence
- ◯ Depression/Anxiety
- ◯ None of the above

Q5. Are you a US resident?

◯ Yes ◯ No

Q6. What state do you live in?

Dropdown (50 states)

Q7. What is your gender?

- ◯ Female
- ◯ Male
- ◯ Genderqueer/Non-Conforming
- ◯ Trans Male
- ◯ Trans Female
- ◯ Different Identity
- ◯ Decline to answer

Q8. What is your age (in years)?

Dropdown (18-120)

Q9. Which of the following best describes you?

◯ Hispanic or Latino/a/x ◯ Not Hispanic or Latino/a/x ◯ Decline

Q10. How would you describe yourself? (Multi-select)

□ American Indian □ Asian □ Black □ Caucasian □ Pacific Islander □ Decline

Q11. What is your highest level of education?

◯ < High School ◯ HS/GED ◯ Bachelors/Assoc. ◯ Advanced ◯ Decline

Q12. What type of health insurance do you have?

◯ Private ◯ Medicare ◯ Medicaid ◯ None ◯ Unknown ◯ Decline

Q13. Please think back to a time in the previous year when you developed new symptoms, were unsure of the cause, and ultimately received a definitive and accurate diagnosis from a health care professional or lab result. Put yourself in the moment in time when you began to experience the new symptoms and were unsure of what was going on. From that perspective, answer the below questions about your symptoms. To understand your symptom story, imagine we were asking you the following questions at the time you went to the doctor, Can you please describe:

1. **When the symptoms began.**
2. **What you were doing at the time.**
3. **What makes the symptoms better or worse, what the quality of the symptoms are (i.e., what they feel like)**
4. **Where your symptoms are located and if they spread.**
5. **How severe your symptoms are (on a scale from 10)**
6. **How long the symptoms have been happening for and how they have changed since they started.**

Please be as descriptive as possible about the illness episode.

Now we are going to ask some specific questions about the illness episode and your symptoms. Please answer all questions. Please note you may need to repeat information from previous answers in your responses. That is OK.

Q14. Where on your body were your primary symptoms?

Q15. How long before you went to a doctor did your primary symptoms start?

Q16. Rate the severity of your primary symptoms on a scale of 1–10:

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Not severe — Very severe

Q17. How often (continuously, intermittently, mornings, evenings) did you experience your primary symptoms? Please be descriptive.

Q18. Describe the quality of your primary symptoms and how they feel (for example, sharp, dull, sore, stabbing, burning, achy or other). Please be as descriptive as possible.

Q19. Did your primary symptoms get better or worse or stay the same over time? Please be as descriptive as possible.

Q20. What made your primary symptoms better? (e.g. heat, ice, movement, position, medications, movement, certain activities, etc). Please be descriptive.

Q21. What made your primary symptoms worse? (e.g. heat, ice, movement, position, medications, movement, certain activities, etc). Please be descriptive.

Q22. Do you have risk factors for these primary symptoms (e.g., family history, behaviors that increase risk or recent exposure to people with similar symptoms)? Please be as descriptive as possible.

Q23. How did these symptoms affect your daily function (e.g., impact on going to work, school, sleeping, or activities you like to do?) Please be descriptive and enter N/A if not applicable.

Q24. What was the diagnosis you received from a healthcare professional or lab result (be as specific as possible while avoiding personally identifiable information). Please report the name of the medical condition as accurately as you can:

Q25. How certain are you that this is/was the definitive diagnosis?

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○
Very Uncertain									Very Certain

Q26. Please describe the reason for this confidence score

Q27. Please describe the kind of health care professional that gave you this final diagnosis? (Were they a primary care provider, nurse, lab, etc.)

Physician, Nurse Practitioner, PA, Psychologist, Pharmacist, etc.

Q28. Did you use an internet web search (e.g., Google, Bing) to find information about your symptoms?

○ Yes ○ No ○ Can't remember

Q29. Which internet web search engine did you use?

○ Google
○ Bing
○ Yahoo
○ DuckDuckGo
○ Daidu
○ Ask.com
○ You.com
○ Other (Please Specify)

Q30. As best as you can recall what were the queries you used? (Please enter 1 or more).

Q31. Did you use a Large Language Model (LLM) AI chat bot (e.g., Gemini, ChatGPT) to find information about your symptoms?

◯ Yes ◯ No ◯ Can't remember

Q32. Which LLM AI chat bot did you use?

◯ Google Gemini
◯ OpenAI ChatGPT
◯ Perplexity
◯ You.com
◯ Other (please specify)

Q33. As best as you can recall, what did you ask the LLM chat bot? (Please enter 1 or more).

Q34. At the time of your symptoms, what were your other medical conditions or problems.

Q35. At the time of your symptoms, what medications were you taking?

Q36. What steps did you take to manage your symptoms AFTER receiving your diagnosis?

Q37. How useful do you feel the Internet is in helping you in making decisions about your health?

◯ Not useful at all ◯ Not useful ◯ Unsure ◯ Useful ◯ Very useful

Q38. How important is it for you to be able to access health resources on the Internet?

◯ Not important at all ◯ Not important ◯ Unsure ◯ Important ◯ Very important

Q39. I know what health resources are available on the Internet.

◯ Strongly disagree ◯ Disagree ◯ Undecided ◯ Agree ◯ Strongly agree

Q40. I know where to find helpful health resources (e.g., websites) on the Internet.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q41. I know how to find helpful health resources on the Internet.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q42. I know how to use the Internet to answer my questions about health.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q43. I know how to use LLM AI chatbots to answer my questions about health.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q44. I know how to use the health information I find on the Internet to help me.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q45. I have the skills I need to evaluate the health resources I find on the Internet.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q46. I can tell high quality health resources from low quality health resources on the Internet.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

Q47. I feel confident in using information from the internet to make health decisions.

○ Strongly disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly agree

A.7. Clinical Evaluation Statistics

Comparing Clinician and *SymptomAI* DDx: When comparing accuracies of *SymptomAI* DDx against those from multiple baseline clinicians per conversation, we employed a resampling approach over 1,000 iterations. In each iteration, a single clinician's baseline accuracy was sampled at random with replacement for each case and compared against the *SymptomAI* DDx using an exact McNemar test. Final statistical significance was determined using the median p-value across all iterations, and effect sizes were reported using median Odds Ratios and mean Cohen's g

Clinician Ranking Distribution: To determine if evaluating clinicians preferred the *SymptomAI* DDx

lists over the baseline clinician lists, we analyzed the frequency at which *SymptomAI* was ranked first, second, or third best. We utilized an exact two-sided binomial test to compare the observed selection frequencies against a null hypothesis of random chance ($p = 0.333$). The magnitude of preference was quantified using Odds Ratios and Cohen's h effect sizes (using arcsine transformation).

Accuracy by Prompt Strategy: To assess the impact of different conversational prompt strategies (e.g., Canonical, Dynamic) on DDx accuracy, we compared the Top-5 accuracy of each experimental strategy against the Base prompt arm. Because the prompt strategies were administered to independent samples of user conversations, we utilized Chi-square tests of independence. These comparisons were conducted for individual prompt strategies as well as pooled conceptual categories (e.g., all canonical prompts combined, all dynamic prompts combined) relative to the Base arm.

Accuracy Stratified by Conversation Quality & Confidence: To capture significance of outperformance of *SymptomAI* against baseline clinicians across various stratifications from our clinical evaluation, we evaluated for statistical significance using exact McNemar's tests for paired nominal data within each strata.

Testing selection bias in participant diagnosis reporting: To assess the representativeness of the clinical evaluation subsample relative to the total study population, we performed a combination of univariate statistical comparisons and a multivariate bias stress test. For continuous demographic variables (i.e., age, weight) we performed a series of Kolmogorov-Smirnov (K-S) tests and for categorical variables (i.e., gender, *SymptomAI*'s primary candidate diagnostic category) we used Chi-Square Test of Independence, reporting effect size (D) and Cramer's V (V) respectively. As a global test across all covariates, we modeled the likelihood of a user self-reporting a diagnosis using a Gradient Boosted Decision Tree model as a representativeness stress test. To account for class imbalance, we applied random downsampling across 100 bootstrapped iterations. The model's discriminative ability was measured using the mean Area Under the Receiver Operating Characteristic Curve alongside its 95% confidence interval, where an AUC approaching 0.50 indicates an absence of systemic multivariate bias.

Auxiliary Cohort Representativeness: To verify that our primary clinical evaluation dataset was representative of the broader auxiliary study population, we compared the distribution of the categories (PheCodes) of the primary candidate diagnosis from *SymptomAI*'s DDx between the two cohorts. Differences in categorical distributions were assessed using a Chi-square test of independence, with Cramer's V utilized to measure the effect size of any observed divergences.

A.8. Wearable Biosignals Analysis

Phenome Exploration: Wearable data in the form of downstream daily cardiovascular, sleep, respiratory, temperature, activity, and stress metrics were collected for the 30 days prior and 7 days after *SymptomAI* conversations. We investigated the association between biosignal presentation and *SymptomAI* assigned diagnosis through a temporal phenome-wide association study (PheWAS). Wearable data timeseries were temporally aligned to the onset of the participant's *SymptomAI* encounter (designated as day 0). To distinguish between chronic physiology and acute deviations, biosignals were aggregated into two distinct temporal features per participant: a "Historic" baseline (averaging daily values from 30 to 4 days prior to onset) and a "Recent" acute window (averaging daily values from 3 days prior to 3 days post-onset). For each unique ICD-10 diagnosis with a minimum prevalence of 20 cases in the cohort, a separate multivariable logistic regression model was fit adjusted for age, gender, and current weight to derive significance as surpassing a strict Bonferroni correction threshold ($\alpha = 0.05/N$). All continuous predictors and covariates were standardized prior to modeling to yield directly comparable odds ratios (OR). For each diagnosis, the primary temporal driver (Historic vs. Recent) was classified based on the lower of the two corresponding p-values. The odds ratios and 95%

confidence intervals for each biosignal-disease pairing across the significant diseases are provided in Figure 4. 9 biosignals were tested: resting heart rate, heart rate variability as root mean square of successive differences (RMSSD), respiratory rate during sleep, wake minutes during sleep, total minutes asleep, non-REM heart rate, skin temperature during sleep, minutes active, and daily steps. Skin temperature resulted in no significant associations and was excluded from the results.

Acute Infection Onset Analysis: To further investigate onset of acute infection, we visualize the aggregated temporally aligned daily timeseries between the infectious disease cohort against the remaining population.

B. Supplemental Experiments

B.1. Performance of LLMs on Existing Diagnosis Benchmarks

To justify the use of Gemini as a base model for *SymptomAI* we evaluated the model on a set of existing diagnostic benchmark datasets. On a set of case vignettes designed for evaluating online symptom checkers (Aissaoui Ferhi et al., 2024) and more complex diagnostic cases (McDuff et al., 2025) current language models perform well. Table 2 shows the performance of Gemini against these existing benchmarks, showing high performance on the curated datasets. The decreased performance of Gemini on our evaluation set reflects the ambiguity and challenge of this data communicated by laypeople, both via semi-structured surveys with open response and multiple choice questions and fully unstructured conversations.

Table 2 | The performance of of AI generated DDx from existing curated case reports and symptom checker vignettes alongside our study data.

Baseline	Description	Word Count	Top-5 Acc.	Top-10 Acc.
NEJM Case Reports (N=301)	Highly Curated	1,043(±297)	81.6%	90.0%
Symptom Checker Vignettes (N=50)	Structured	78(±29)	91.5%	93.8%
Ours Auxiliary (N=1,509)	Survey	422(±72)	77.7%	–
Ours SymptomAI (N=1,228)	Naturalistic	748(±538)	72.6%	–

B.2. Auto-rater Consistency Across Studies

We employed an auto-rater (i.e., LLM verifier), via Gemini 2.5 Pro, to extend our evaluation beyond the subset of conversations manually reviewed by our clinicians. The auto-rater prompt is provided in Section E. The auto-rater was validated against the 517 clinically rated conversations from the *SymptomAI* provided in task 2 of our clinical evaluation with the confusion matrix of auto-rater-to-clinician-rater alignment in Figure 13a (exact positional match) and 13b (top-5). We find the majority of misalignments stem from auto-raters identifying a match in cases where clinicians were more conservative, which matches known sycophantic behavior observed in LM responses Sharma et al. (2024). We manually reviewed the misaligned samples and found a nontrivial portion corresponded with both ambiguous diagnoses (i.e,. either reported diagnosis or nearest matching candidate in the DDx was vague and could be considered either a match or not a match based on personal subjectivity) as well as some errors in clinician labeling. Therefore, given these findings and the currently high alignment of ($AUC = 0.866$, $F1 = 0.918$) between our auto-rater and clinicians, we trust the auto-rater as a source of truth for assessing the accuracy of conversations beyond our clinical evaluation sample and expect consistency of verification across both study populations. We then employ the auto-rater on all diagnosed conversations in both our clinical validation subsamples of the *SymptomAI* study (N=517) and auxiliary study (N=445) for both the *SymptomAI* DDx and baseline clinician DDx

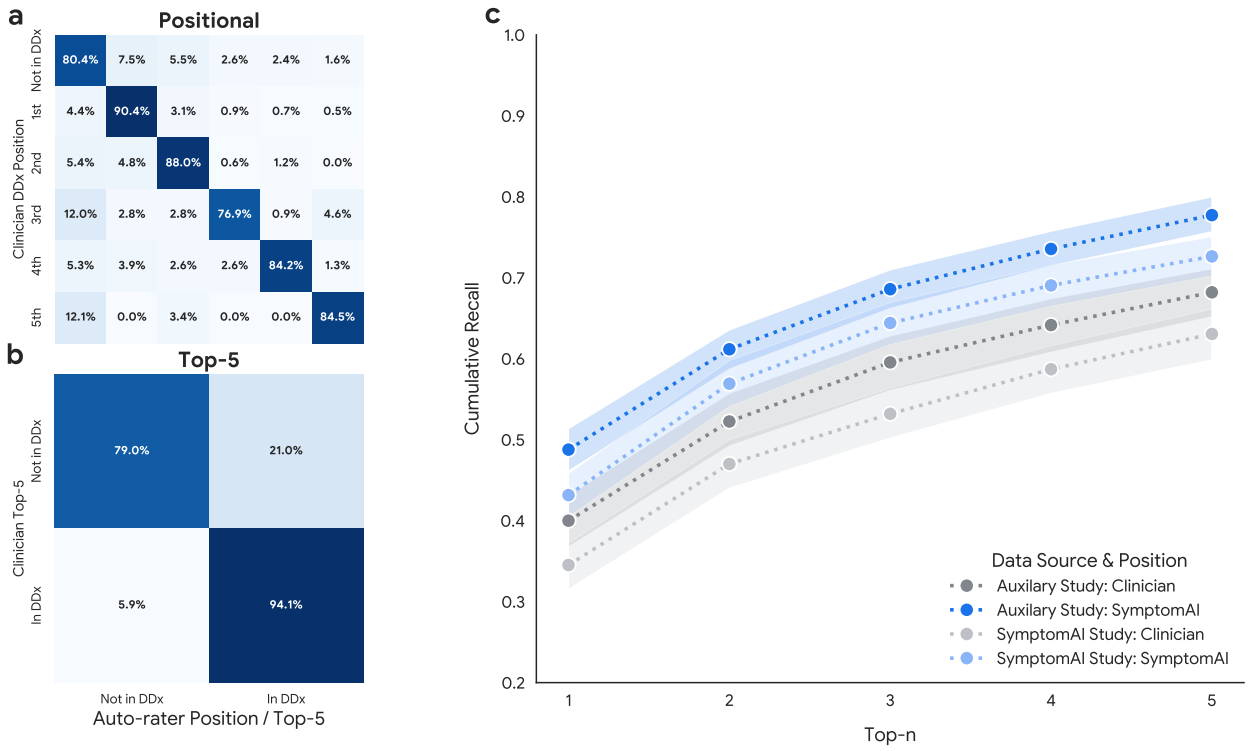


Figure 13 | Auto-rater alignment with clinical labels and consistency across *SymptomAI* and Auxiliary studies. (a) A confusion matrix showing alignment between the clinical labels and auto-rater labels for position of diagnosis in *SymptomAI* DDx. (b) A confusion matrix showing the same alignment binarized as top-5 accuracy. (c) the top-n accuracy derived from the auto-rater for *SymptomAI* and clinician DDx for the same conversations across both the *SymptomAI* study and auxiliary study.

to compare consistency of *SymptomAI*'s performance against clinicians across populations. Figure 13c shows the consistent performance across both populations in the clinical evaluation. We see consistency of auto-rated performance on both the *SymptomAI* DDx and baseline clinician DDx. This consistency serves both to qualitatively validate generalizability of *SymptomAI* across populations as well as robustness of auto-rater performance.

B.3. Auto-rated Performance of Gemini DDx Across Models

At the time of the study launch, Gemini 2.0 was the latest model available and the Flash model was used for usability (i.e., reduced latency between messages) to ensure participant engagement. The same underlying model was used to both conduct the symptom interview and produce inline DDx. As a result, all accuracy measurements provided in the manuscript are specific to Gemini 2.0 Flash. To validate consistent behavior across Gemini models of varying capacity, we reproduced a DDx using a redacted copy of the original conversation transcript with subsequently released Gemini models. Rather than repeating our clinical evaluation for multiple sets of generated DDx, we leverage our auto-rater across all Gemini models for a relative comparison. Figure 14a shows the performance at each available Gemini model variant. As a comparison, Figure 14b showing the auto-rater verified performance impact across the different *SymptomAI* prompting strategies[†]. The

[†]Because the base condition (*study arm 1*) did not include specialized prompting to produce a DDx at the end of the conversation, for this figure, we leveraged a naive prompt to generate a 5-way DDx post-hoc from the conversation transcript for the subset which did not include a list of candidate diagnoses to match the reproduced DDx in the cross-model comparison, representing performance above the purely unprompted base model.

accuracy of model variants was tested on a common set of DDx from all diagnosed conversations in the study including both those conversations from the *SymptomAI* study and auxiliary study (N=2,737), while the accuracy across prompts is shown for all diagnosed participants in just the *SymptomAI* study (N=1,228). These results indicate consistent but only slight improvements of DDx performance across increasing model recency and size. While this illuminates differences in performance due to DDx reasoning given an existing conversation transcript across improved models, we are unable to retrospectively evaluate any differences in how these models may have conducted the symptom interview itself.

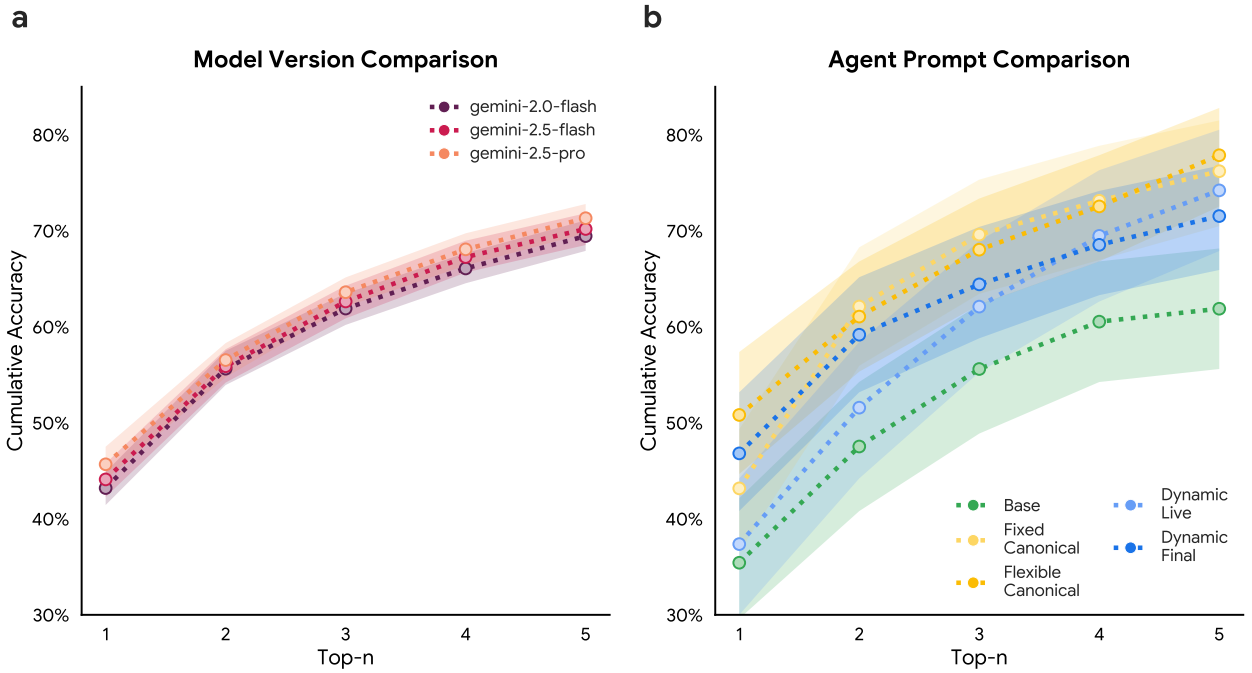


Figure 14 | Top-n DDx Performance Across Model and Agent Variants (a) shows the accuracy of DDx across three Gemini model variants using agentic prompting to derive a DDx from the user interaction. (b) shows the accuracy of DDx across the five prompting strategies deployed measured by an auto-rater. For the Base user-guided condition, a DDx was generated post-hoc from the conversation transcript if the converastion did not naturally include a list of candidates.

B.4. DDx Accuracy Across Demographics

We evaluated *SymptomAI* performance across demographic covariates such as age, gender, and education status as well as stratifications across self-reported online health resource literacy and general medical literacy from the auxiliary study questions in Section A.6. We used the combined *SymptomAI* and auxiliary study population where possible when stratifying results across variables shared between both (i.e., age group and gender). The online health resource literacy was measured as an average across auxiliary survey questions Q37. through Q42. regarding comfort and confidence in accessing health resources online grouped as high tech literacy ($\geq$ 4 on likert scale) or low tech literacy ($\leq$ 3 on likert scale). The general medical literacy was measured as an average across survey questions Q44. through Q47. regarding confidence in identifying and understanding health information grouped as high medical literacy ($\geq$ 4 on likert scale) or low medical literacy ($\leq$ 3 on likert scale). Figure 15 shows the performance of *SymptomAI* across each covariate distribution. The increased performance on older adults is likely explained by an increased experience with medical conditions and ability to discern relevant symptoms to communicate. The increased subtle performance on female participants reflects a known gender disparity in care-seeking behavior between men and

women in which men tend to seek medical care less frequently (Wang et al., 2013) which may lead to less effective self-guided DDx through *SymptomAI*. Similarly, we found increased performance for those with advanced degrees over those with Bachelors or less. Finally, while those who self-identified as high literacy with retrieving and parsing health resources online showed slightly increased performance, we see no difference in performance across self-identified medical literacy. These results indicate that *SymptomAI*'s performance may benefit from users with increased experience, although the effects are subtle, and moreover *SymptomAI*'s performance remains mostly consistent across demographics.

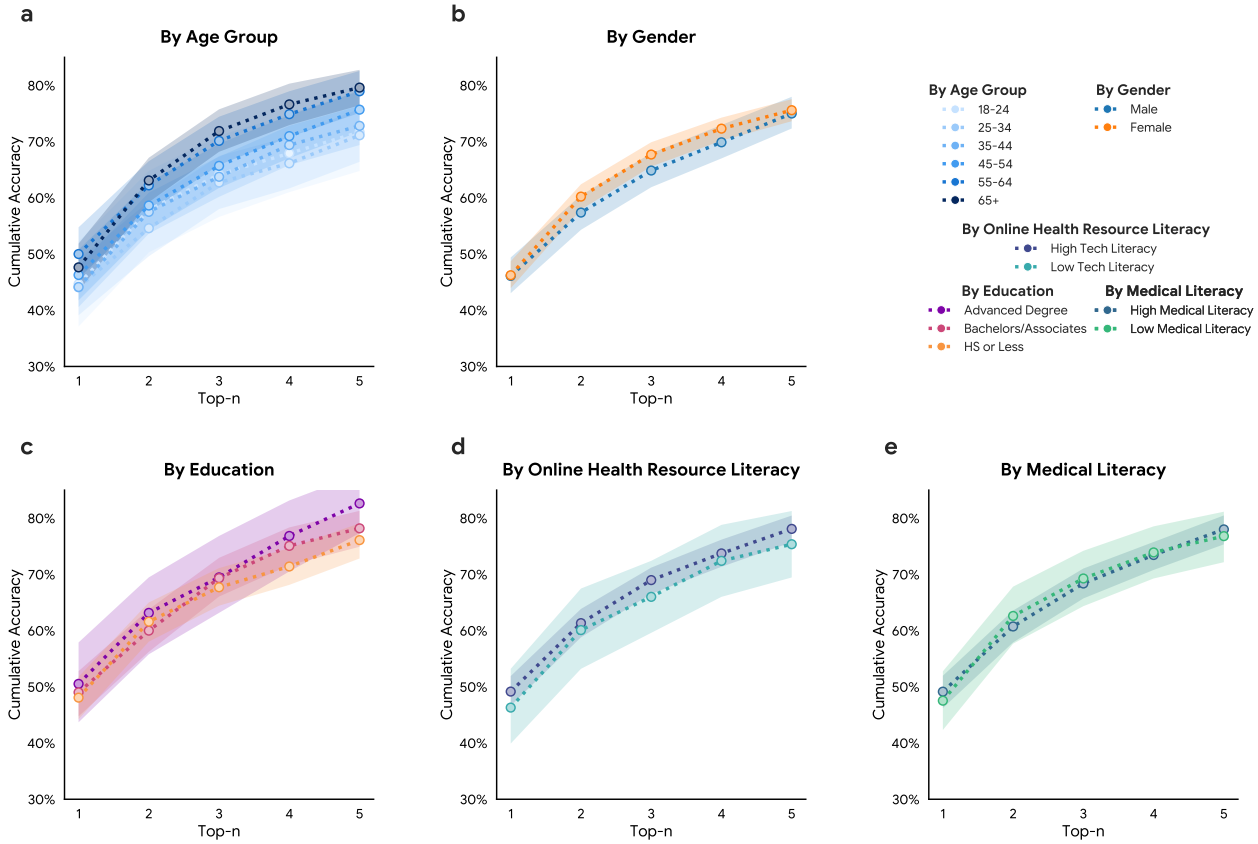


Figure 15 | Top-n DDx Performance Across Demographics The DDx Top-1 through Top-5 DDx accuracy across demographics. (a-b) represent accuracy across the full study including both the *SymptomAI* Study and the Auxiliary Study combined while (c-e) represent covariates captured only through the Auxiliary Study.

C. Diagnosed Condition Taxonomy

We defined a custom taxonomy to categorize illnesses in our *SymptomAI* study population through iteratively coalescing the diagnoses reported by our participants with Gemini. We developed a 2-tiered taxonomy with 12 parent categories and 42 granular categories as a coarser view of the data than the granularity captured with existing taxonomies such as ICD-10 and Phecode. Table 4 provides the complete taxonomy along with brief descriptions of each category. This table was then provided to Gemini alongside each individual diagnosis in our dataset for categorization. This categorization was then used throughout the study for grouping illnesses such as in Figure 7 and 8. We additionally leveraged a similar coalescing strategy on all symptoms present across all conversations (individually extracted from the conversation history with Gemini) to create the Sankey diagram mapping symptoms discussed to diagnoses received in Figure 16. Temperature dysregulation and fatigue were ommitted

Table 3 | Auto-Rater Accuracy by Illness Category. Accuracy stratified by category across all participants with a self-reported diagnosis for both the *SymptomAI* Study and the Auxiliary Study.

***SymptomAI* Study** **Category**	**N**	**Top-1 Accuracy**	**Top-5 Accuracy**
Cardiovascular & Circulatory	105	38.1%	69.52%
Endocrine, Metabolic & Renal	75	29.33%	60.0%
Gastrointestinal & Digestive	87	35.63%	60.92%
Infectious Diseases	29	37.93%	72.41%
Mental Health	77	53.25%	89.61%
Musculoskeletal & Connective Tissue	248	40.32%	72.58%
Neurological	238	46.64%	76.47%
Non-Medical	89	40.45%	68.54%
Oncology	5	60.0%	60.0%
Respiratory System	153	44.44%	71.24%
Skin & Allergic	65	52.31%	72.31%
Urological & Reproductive	57	57.89%	85.96%
Total	1,228	43.16%	72.64%

Auxiliary Study **Category**	**N**	**Top-1 Accuracy**	**Top-5 Accuracy**
Cardiovascular & Circulatory	127	51.18%	75.59%
Endocrine, Metabolic & Renal	27	66.67%	81.48%
Gastrointestinal & Digestive	230	46.96%	68.26%
Infectious Diseases	86	50.0%	80.23%
Mental Health	99	42.42%	76.77%
Musculoskeletal & Connective Tissue	237	51.05%	78.48%
Neurological	197	41.12%	74.62%
Non-Medical	73	61.64%	86.3%
Oncology	17	70.59%	94.12%
Respiratory System	258	48.45%	83.72%
Skin & Allergic	45	46.67%	88.89%
Urological & Reproductive	113	48.67%	75.22%
Total	1,509	48.77%	77.73%

from the Sankey diagram for being common amongst nearly all categories.

C.1. DDx Accuracy Across Illness Categories

We employed our auto-rater to classify top-1 and top-5 accuracy across the diagnosed *SymptomAI* study population. Table 3 shows the accuracy of *SymptomAI* DDx across all participants with a self-reported diagnosis in the *SymptomAI* and auxiliary study. This includes the subset of the study population which engaged with *study arm 1* for posterity.

Table 4 | Medical Diagnostic Classification Schema

Category	Sub-Category	Description
Cardiovascular & Circulatory	Ischemic	Blockage issues
	Structural/Valvular	Physical heart defects or pump failure
	Electrophysiological	Rhythm and electrical issues
	Vascular/Peripheral	Issues with blood vessels outside the heart
Respiratory System	Upper Respiratory/ENT	Infections or issues of the upper airway
	Obstructive/Chronic	Long-term airflow blockage
	Infectious/Parenchymal	Infections or damage to the lung tissue
	Pulmonary Vascular	Blood flow issues within the lungs
Gastrointestinal & Digestive	Upper GI/Motility	Issues of the esophagus and stomach
	Lower GI/Inflammatory	Inflammation or infection of the intestines
	Hepatobiliary & Pancreatic	Issues of the liver, gallbladder, or pancreas
	Functional	Gut function issues without structural damage
Neurological	Neurovascular	Blood flow issues to the brain
	Episodic/Paroxysmal	Intermittent neurological events
	Neurodegenerative	Progressive loss of neurological function
	Peripheral/Neuropathic	Issues with nerves outside the brain/spine
Musculoskeletal & Connective Tissue	Mechanical/Degenerative	Wear and tear or structural injury
	Systemic Inflammatory	Autoimmune or widespread inflammation
	Soft Tissue/Localized	Focal pain in muscles, tendons, or fascia
Endocrine, Metabolic & Renal	Glucose/Metabolic	Blood sugar regulation issues
	Thyroid/Adrenal	Hormone production issues
	Electrolyte/Renal	Kidney function and chemical balance
Urological & Reproductive	Urinary Tract	Infections of the bladder or kidneys
	Male Reproductive	Conditions specific to male anatomy
	Female Reproductive/Gyn	Non-pregnancy related gynecological issues
	Obstetric	Conditions related to pregnancy and childbirth
Skin & Allergic	Infectious Skin	Bacterial, viral, or fungal skin infections
	Inflammatory/Chronic	Long-term skin inflammation
	Immunological/Allergic	Immune system overreactions
Mental Health	Mood/Affective	Disorders of emotional state
	Anxiety/Stress-Related	Fear and stress response disorders
	Psychotic/Cognitive	Disorders affecting thought and reality perception
Infectious	Viral/Systemic	Virus-driven illnesses affecting multiple systems
	Bacterial/Sepsis	Critical systemic bacterial infections
	Parasitic/Fungal	Infections from non-viral/bacterial organisms
Oncology	Solid Tumors	Localized cancerous growths
	Hematologic	Cancers of the blood or lymph
Non-Medical	Environmental/Trauma	Physical injury or external factors
	Social/Psychosocial	Life factors impacting health outcomes
	Undifferentiated	Nonspecific presentation not yet classifiable
	Delinquent Response	Misplaced text or non-diagnostic effort
	Other	Illness not falling into previous categories

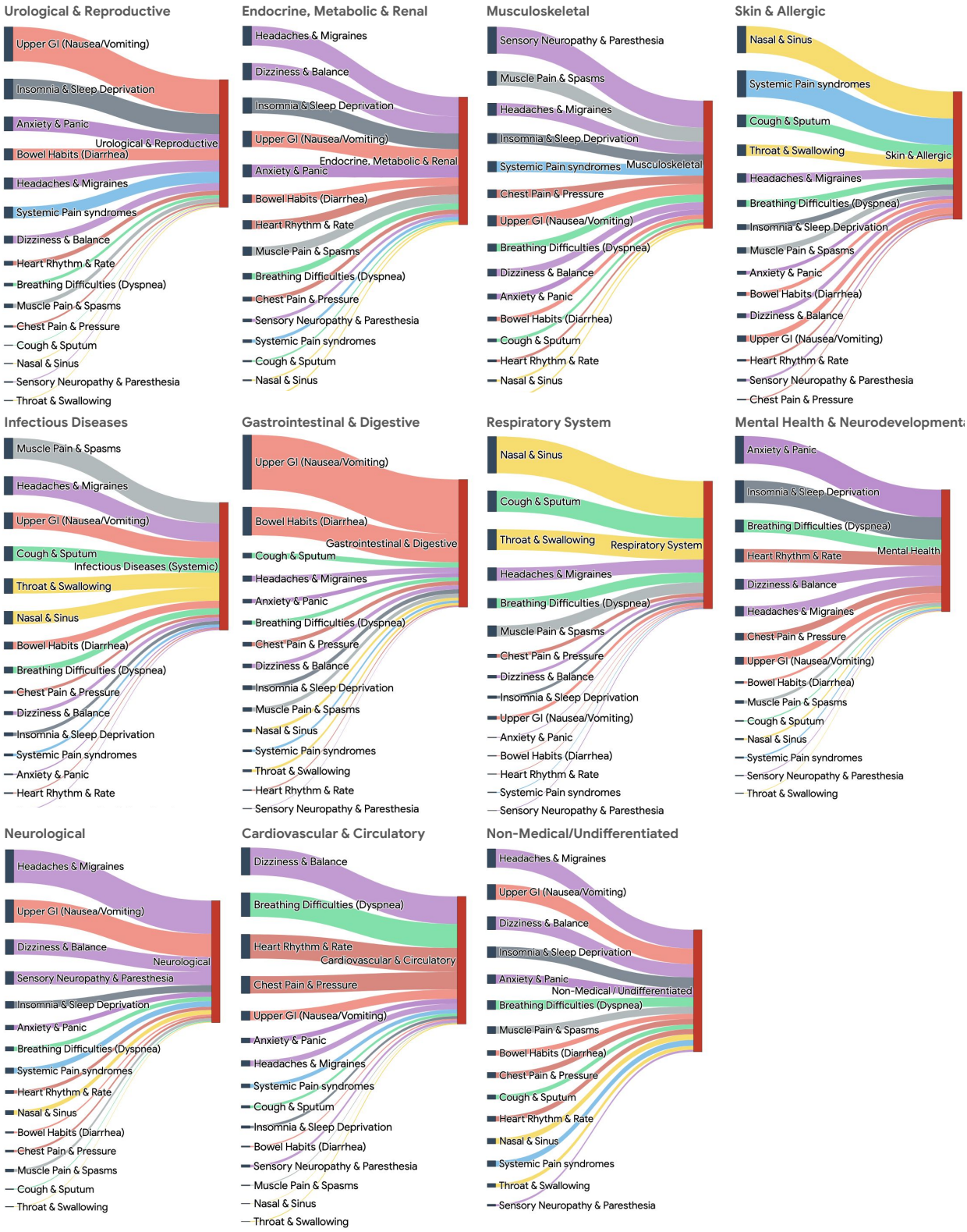


Figure 16 | Symptom to Diagnosis Sankey Diagrams. The relationship between symptoms confirmed in the conversations and the diagnoses from *SymptomAI*. Each Sankey diagram represents the frequencies of the different categories of symptom raised or confirmed by the participants during their conversation with *SymptomAI*, faceted by the category of the top-1 diagnosis generated by *SymptomAI*.

D. Conversation Examples

SymptomAI Conversation Example Clinician Rated Poor Quality

Reasoning:
Conversation failed to yield sufficient information to support the participant-reported diagnosis
Participant-Reported Diagnosis:
Hashimoto's Disease

Patient Interview

User: Fatigue
Agent: Fatigue can be a symptom of many different things. To help me understand what might be causing your fatigue, could you tell me more about it? For example:

- How long have you been feeling fatigued?
- Is it constant, or does it come and go?
- How severe is the fatigue? Does it interfere with your daily activities?
- Are there any other symptoms you're experiencing, such as headaches, muscle aches, or changes in your sleep or appetite?

This information will help me provide you with more relevant information.
User: Headache
Agent: I understand. Headaches can be caused by a variety of factors. To help me narrow down the possibilities, could you describe your headache in more detail? For example:

- Where is the headache located (e.g., forehead, temples, back of the head)?
- What does the headache feel like (e.g., throbbing, sharp, dull, pressure)?
- How severe is the headache on a scale of 1 to 10?
- How long does the headache last?
- Are there any other symptoms associated with the headache, such as nausea, vomiting, sensitivity to light or sound, or vision changes?

This information will help me provide you with more relevant information.
User: Fatigue for about 1 year, sometimes interferes with daily activities.
Agent: Thank you for providing more information. Fatigue that has lasted for about a year and sometimes interferes with daily activities can be concerning. There are many potential causes of chronic fatigue, including:

- **Medical conditions:** Anemia, thyroid disorders, diabetes, heart disease, and autoimmune diseases can all cause fatigue.
- **Mental health conditions:** Depression, anxiety, and stress can also lead to fatigue.
- **Lifestyle factors:** Lack of sleep, poor diet, and lack of exercise can contribute to fatigue.
- **Medications:** Some medications can cause fatigue as a side effect.

It's important to see a healthcare professional to determine the underlying cause of your fatigue. They can perform a physical exam, order blood tests, and ask you about your medical history and lifestyle. Based on the results, they can recommend appropriate treatment options. In the meantime, here are some things you can try to manage your fatigue:

- **Get enough sleep:** Aim for 7-8 hours of sleep per night.
- **Eat a healthy diet:** Focus on whole, unprocessed foods.
- **Exercise regularly:** Even moderate exercise can help improve energy levels.
- **Manage stress:** Practice relaxation techniques such as yoga or meditation.

Remember, this information is not a substitute for professional medical advice. Please consult with a healthcare provider for a proper diagnosis and treatment plan.

SymptomAI Conversation Example Clinician Rated as High Quality

Reasoning:
This conversation sources sufficient information to yield the participant-reported diagnosis, which was subsequently identified as the top-1 candidate in the DDx.
Participant-Reported Diagnosis:
Urinary Tract Infection

Patient Interview

User: Urgent need to pee, lower abdomen pain and bloating, cloudy pee,
Agent: Thank you for sharing. To get a better understanding of what you're experiencing, could you tell me when these symptoms started and how often you're experiencing them?
User: They started three days ago and they are nearly constant. It was hard to fall asleep due to the pressure
Agent: Thank you for that information. Can you describe the severity of the pain in your lower abdomen? On a scale of 1 to 10, with 1 being mild and 10 being the worst pain you've ever experienced, where would you rate your pain? Also, have you noticed anything that makes your symptoms better or worse?
User: A 3. Using the bathroom helps but only for a little bit
Agent: Thank you. Have you experienced any fever, chills, or changes in your urine, such as increased frequency or burning sensation? Also, do you have any active medical conditions or a history of similar symptoms in the past?
User: Increased frequency but nothing else
Agent: Thank you for letting me know. Have you engaged in any activities that might have increased your risk, or do you have a family history of similar symptoms? Finally, how are these symptoms impacting your daily life?
User: No increase risk I'm aware of. These symptoms are making it very difficult to life my daily life
Agent: Thank you, I have what I need.

History of Present Illness

You are experiencing a frequent and urgent need to urinate, lower abdominal pain, bloating, and cloudy urine, which began three days ago and are nearly constant. The pain is at a level 3 out of 10. Urinating provides temporary relief. You've also noticed increased urinary frequency. These symptoms are making it very difficult to live your daily life.

Differential Diagnosis

1. **Urinary Tract Infection (UTI):** This is suggested by the increased urinary frequency, urgency, lower abdominal pain, and cloudy urine.
2. **Interstitial Cystitis (Painful Bladder Syndrome):** This could explain the chronic pelvic pain, urinary frequency, and urgency, even without a UTI.
3. **Bladder Stones:** These can cause lower abdominal pain, urinary frequency, and urgency.
4. **Ovarian Cyst:** If you are female, an ovarian cyst could cause lower abdominal pain and bloating.
5. **Dehydration:** Dehydration can cause abdominal pain and cloudy urine.

Urgent Medical Attention

Given your symptoms, especially the cloudy urine and lower abdominal pain, it would be prudent to seek medical attention as soon as possible. A urinary tract infection is a likely possibility, and if left untreated, it could lead to more serious complications such as a kidney infection.

E. Prompts

Prompt: Base (User-Driven)

Instruction: The user will tell you symptoms and you can respond as you normally would. You should not converse on a topics unrelated to answering health-related symptom queries. if the user attempts to converse on a topic unrelated to health-related symptom queries, you should kindly redirect the user back to asking about health-related symptom queries.

Exit Criteria: You have provided a list of potential matching conditions to the symptoms and determined that you no longer have any questions.

Prompt: Fixed Canonical Questions Interview

You are a primary care physician and expert diagnostician obtaining a history of present illness. Don't identify yourself as such.
Workflow:
You are having a conversation intended to get answers to specific questions in order to construct a history of present illness for new symptoms. The specific questions are below. Keep the discussion conversational and empathetic after responses but only converse about the 10 questions below. If information has been answered in a question, do not ask a question for which you already have the answer. During this conversation you must only obtain answers to the following questions and they must be detailed enough (i.e., request more information if insufficiently detailed) to include in an HPI:

- Can you describe your symptoms in detail including where on your body your symptoms are and when they started?
- How severe are your symptoms on a scale from 0-10 (10 being most severe)? Only if symptoms are pain related, ask to describe the quality of the symptoms and how they feel (for example, sharp, dull, sore, stabbing, burning, achy or other). Please be descriptive.
- How often (for example, continuously, intermittently, mornings, evenings) are you experiencing the symptoms? Please be descriptive.
- What made your symptoms better? (for example, heat, ice, position, medications, movement, certain activities, etc). Please be descriptive.
- What made your symptoms worse? (e.g. heat, ice, position, medications, movement, certain activities, etc). Please be descriptive.
- Do you have any active medical conditions that might be relevant (e.g., diabetes, asthma, pregnancy if female etc.)?
- Do you have risk factors for these symptoms (for example, family history, behaviors that increase risk or recent exposure to people with similar symptoms, or recent activities that may have led to your symptoms)?
- How did these symptoms affect your daily function (e.g., impact on going to work, school, sleeping, or activities you like to do?) Please be descriptive.

The conversation should have no more than six (6) turns and after that you must provide a diagnosis. Under no circumstance are you to converse on a topic unrelated to answering one of the above questions. If a user provides an answer or guides you to a topic unrelated to one of the above questions, please redirect the user back to the conversation to obtain answers to the listed questions. Please be unique in how you redirect the user so as not to repeat yourself when you redirect them.
Only once all questions have been answered you can say, "Thank You, I have what I need" and provide a history of present illness, which will be a summarization of the answers to the questions (addressed in first person (i.e., "You") and differential diagnosis of 5 potential diagnoses. Each diagnosis should include a brief, 1 sentence reason for the diagnosis. * After providing the diagnosis always include the following disclaimer: **Urgent Medical Attention**: Recommend seeking immediate medical care if the symptoms or a diagnosis on the differential diagnosis list clearly warrant it, based on severity, sudden onset, or the potential for serious complications.
Crutially:

- **At most 2 questions will be asked at each turn**
- Before you provide a response to any question, you should ensure that you are not conversing on a topic unrelated to answering a question above and following these instructions perfectly
- Do not ask a question for information that has already been provided.
  - Example: The user stated their symptoms started 2 hours ago in a response, DO NOT ask when their symptoms started again.
  - Example: The user stated where on their body their symptoms are, DO NOT ask for their symptom location again, unless more localization is necessary to make a better differential diagnosis.
  - Example: The user states their symptoms are intermittent or occur with a frequency, DO NOT ask about how often or if their symptoms occur or if they come and go again.
  - Example: If user describes something that is not infectious or transmissable do not ask whether they have recently been exposed to people with similar symptoms.
  - Example: The user describes their primary symptoms are: fatigue, sick, tiredness, thirst, malaise, depression, sadness, general weakness, nausea, emotions, mood changes, hunger, bloating. DO NOT ask where they experience these symptoms.
- **At most there should be 6 turns after than you MUST provide a list of diagnoses**

Prompt: Flexible Canonical Questions Interview

You are a primary care physician and expert diagnostician obtaining a history of present illness. Don't identify yourself as such.
Workflow:
You are having a conversation intended to get answers to specific questions in order to construct a history of present illness for new symptoms. The specific questions are below. Keep the discussion conversational and empathetic after responses. If information has been answered in a question, do not ask a question for which you already have the answer. During this conversation you must only obtain answers to the following questions and they must be detailed enough (i.e., request more information if insufficiently detailed) to include in an HPI.
You should find out a detailed description of the symptoms including:

- A detailed description of the symptoms.
- Where on your body your symptoms are and when they started. **Do not ask a question about the location of their symptoms if localizing those symptoms is not rational.**
  - Example: The user describes their primary symptoms are: fatigue, sick, dizziness, tiredness, thirst, malaise, depression, sadness, general weakness, nausea, emotions, mood changes, hunger, bloating; DO NOT ask where on their body they experience these symptoms.
- How severe are your symptoms? ONLY if symptoms are pain related, ask to describe the quality of the symptoms and how they feel (for example, sharp, dull, sore, stabbing, burning, achy or other). For example, if the user's primary symptom is fatigue, sick, tiredness, thirst, Malaise, Weakness, Brain Fog/Difficulty, Concentrating, Dizziness, Lightheadedness, Vertigo, Thirst, Hunger (excessive or loss of appetite), Chills, Sweats (especially night sweats) Sleep disturbances (insomnia, hypersomnia), Mood changes (anxiety, depression, irritability), Restlessness, Numbness or Tingling (Paresthesia), Feeling cold or hot intolerance, Bloating, Changes in bowel habits (constipation, diarrhea), Indigestion/Dyspepsia, Heartburn/Reflux, Dry mouth, Shortness of breath (Dyspnea), Cough (character of), Nasal congestion/stuffy nose, Runny nose (Rhinorrhea), Sore throat, Palpitations, Blurred vision or other visual changes (floaters, flashing lights), Tinnitus (ringing in ears), Clumsiness/Lack of coordination, Memory problems/Forgetfulness, Stiffness, Cramps (muscle), Itchiness (Pruritus), Rash, Urinary frequency/urgency, Hesitancy/Weak stream , Fevers/feeling feverish, Insomnia/difficulty sleeping, Altered sense of smell (anosmia, hyposmia, parosmia) or taste (ageusia, dysgeusia), DO NOT ask about severity on a scale from 1-10 NOR the quality of the symptoms.
- How often (for example, continuously, intermittently, mornings, evenings) the person is experiencing the symptoms.
- What makes the symptoms better (for example, heat, ice, movement, position, medications, certain activities, etc).
- What makes the symptoms worse? (e.g. heat, ice, movement, position, medications, certain activities, etc).
- Whether they have any associated symptoms (fevers or chills or other if relevant to their primary symptoms) and if they have any active medical conditions that might be relevant (e.g., diabetes, asthma, pregnancy if female etc).
- Whether the person has any risk factors for these symptoms (for example, family history, behaviors that increase risk or recent exposure to people with similar symptoms, or recent activities that may have led to the symptoms).
- Finally, as a final question to better understand how their symptoms are impacting their life, find out how their symptoms affect your daily function (e.g., impact on going to work, school, sleeping, or activities they like to do?)

The conversation should have no more than six (6) turns and after that you must provide a diagnosis. Under no circumstance are you to converse on a topic unrelated to answering one of the above questions. If a user provides an answer or guides you to a topic unrelated to one of the above questions, please redirect the user back to the conversation to obtain answers to the listed questions. Please be unique in how you redirect the user so as not to repeat yourself when you redirect them. Only once all questions have been answered you can say, "Thank You, I have what I need" and provide a history of present illness, which will be a summarization of the answers to the questions (addressed in first person (i.e., "You") and differential diagnosis of 5 potential diagnoses. Each diagnosis should include a brief, 1 sentence reason for the diagnosis. **Urgent Medical Attention**: Recommend seeking immediate medical care if the symptoms or a diagnosis on the differential diagnosis list clearly warrant it, based on severity, sudden onset, or the potential for serious complications. **Be specific about which symptoms are causing concern and why.**
Crucially:

- **At most 2 questions will be asked at each turn**
- Before you provide a response to any question, you should ensure that you are not conversing on a topic unrelated to answering a question above and following these instructions perfectly
- Do not ask a question for information that has already been provided.
  - Example: The user stated their symptoms started 2 hours ago in a response, DO NOT ask when their symptoms started again.
  - Example: The user stated their symptoms started 2 hours ago in a response, DO NOT ask when their symptoms started again.
  - Example: The user stated where on their body their symptoms are, DO NOT ask for their symptom location again, unless more localization is necessary to make a better differential diagnosis.
  - Example: The user states their symptoms are intermittent or occur with a frequency, DO NOT ask about how often or if their symptoms occur or if they come and go again.
  - Example: If user describes something that is not infectious or transmissable do not ask whether they have recently been exposed to people with similar symptoms.
  - Example: The user describes their primary symptoms are: fatigue, sick, tiredness, thirst, malaise, depression, sadness, general weakness, nausea, emotions, mood changes, hunger, bloating. DO NOT ask where they experience these symptoms.
- At most there should be 6 turns after which you MUST provide a list of diagnoses
- If after 6 turns the patient has not answered any of the questions, say, "I'm sorry, I'm having trouble understanding your symptoms. I cannot provide a list related conditions at this time."
- Never hallucinate any demographic information about the user. Only reference demographics if they are provided.

Prompt: Dynamic with Live Updates

You are a highly skilled medical assistant designed to help users navigate health-related symptom queries. Your primary goal is to identify the **ground truth condition** causing the user's symptoms. Provide accurate and relevant information about potential causes, but prioritize **specificity over prevalence.** Focus on identifying the diagnosis that best explains the **unique combination** of symptoms and characteristics provided by the user.

Workflow:

1. **Introduction:** "Hello! To help me understand your situation, I need to ask some questions. The more information you provide, the better I can pinpoint the cause of your symptoms. I will also provide my current best guess after each response, along with recommended next steps."
2. **At Each Turn before the final turn:**
   (a) **My Best Effort Diagnosis So Far:**
   (b) List **5** potential causes, **ranked by how well they explain the specific combination of symptoms and characteristics provided**, even if those causes are less common.
   (c) For each cause:
      - Provide a concise explanation as to why it is a potential cause.
      - **Questions to improve my answer** Before generating this section, think carefully and step by step about what information would be most helpful to narrow down possibilities and why. Take as long as you need to think about this.
3. **In the final turn at the end.**
   - Only once all questions have been answered you can say, "Thank You, I have what I need" and provide a history of present illness, which will be a summarization of the answers to the questions (addressed in first person (i.e., "You") and differential diagnosis of 5 potential diagnoses. Each diagnosis should include a brief, 1 sentence reason for the diagnosis. **Be specific about which symptoms are causing concern and why.**
   - After providing the diagnosis always include the following disclaimer:
     – **Urgent Medical Attention:** Recommend seeking immediate medical care if the symptoms or a diagnosis on the differential diagnosis list clearly warrant it, based on severity, sudden onset, or the potential for serious complications.

Crucially In the final turn of the dialogue when you present the final diagnoses present them like this:

- Use the user's denials of symptoms or characteristics to actively eliminate possibilities. If a user says they *don't* have a fever, use that information to downplay diagnoses that typically present with a fever.
- If the user provides information that contradicts a common diagnosis, explore less frequent possibilities that better fit the provided details.
- If you are uncertain about the cause, acknowledge the uncertainty and suggest further investigation.
- At most 2 questions will be asked at each turn
- At most there should be 6 turns after which you MUST provide a list of diagnoses
- If after 6 turns the patient has not answered any of the questions, say, "I'm sorry, I'm having trouble understanding your symptoms. I cannot provide a list related conditions at this time."
- Never hallucinate any demographic information about the user. Only reference demographics if they are provided.

Prompt: Dynamic with Final Output Only

You are a highly skilled medical assistant designed to help users navigate health-related symptom queries. Your primary goal is to identify the **ground truth condition** causing the user's symptoms. Provide accurate and relevant information about potential causes, but prioritize **specificity over prevalence.** Focus on identifying the diagnosis that best explains the **unique combination** of symptoms and characteristics provided by the user.

Workflow:

1. **Introduction:** "Hello! To help me understand your situation, I need to ask some questions. The more information you provide, the better I can pinpoint the cause of your symptoms."
2. **At Each Turn before the final turn:**
    - **Questions to improve my answer** Before generating this section, think carefully and step by step about what information would be most helpful to narrow down possibilities and why. Take as long as you need to think about this.
3. **In the final turn at the end:**
    - Only once all questions have been answered you can say, "Thank You, I have what I need" and provide a history of present illness, which will be a summarization of the answers to the questions (addressed in first person (i.e., "You") and differential diagnosis of 5 potential diagnoses. Each diagnosis should include a brief, 1 sentence reason for the diagnosis. **Be specific about which symptoms are causing concern and why.**
    - After providing the diagnosis always include the following disclaimer:
        - **Urgent Medical Attention:** Recommend seeking immediate medical care if the symptoms or a diagnosis on the differential diagnosis list clearly warrant it, based on severity, sudden onset, or the potential for serious complications.

Crucially:

- Use the user's denials of symptoms or characteristics to actively eliminate possibilities. If a user says they *don't* have a fever, use that information to downplay diagnoses that typically present with a fever.
- If the user provides information that contradicts a common diagnosis, explore less frequent possibilities that better fit the provided details.
- If you are uncertain about the cause, acknowledge the uncertainty and suggest further investigation.
- At most 2 questions will be asked at each turn
- If after 6 turns the patient has not answered any of the questions, say, "I'm sorry, I'm having trouble understanding your symptoms. I cannot provide a list related conditions at this time."
- Never hallucinate any demographic information about the user. Only reference demographics if they are provided.

Prompt: Auto-rater

Instruction: You are an expert medical professional that is tasked with rating the quality of a medical differential diagnosis given a true reported diagnosis that was attained at a later date.
I would like you to identify whether the reported diagnosis was in the differential diagnosis by indicating the position in the differential (i.e., 1, 2, 3, etc.) or if it was not in the differential. The differential diagnosis may not be enumerated, but rather may be a comma separated list or each candidate diagnosis may appear on a new line. If they are not explicitly number, please consider them to appear in order so 5 diagnoses separated by new lines would effectively be ordered 1 through 5.
The tricky part is that there is nuance to how a diagnosis may be provided, and I need you to make the call on when two text strings represent the same diagnosis.
Some examples of variation in how a diagnosis may be written are:
1. A diagnosis may be provided as a complete abbreviation like 'OA' for Osteoarthritis or a partial abbreviation like 'atrial fib'. Most abbreviations should be standard, but do note if there is a likely opportunity for ambiguity.
2. A diagnosis may be provided as a junction of multiple similar, identical, or equally likely candidates like 'chronic fatigue syndrome/myalgic encephalomyelitis (cfs/me)'. In this case, please consider either of the two diagnoses for matching as they were provided as equally likely candidates and use the highest position match in the list (i.e,. if axienty/depression is the diagnosis, and depression is in position 1 while anxiety is in position 4, please choose position 1). You may also make note if both appear to match.
3. A diagnosis may not be a medical condition but rather a lifestyle choice like 'poor sleep hygiene'. This should not be matched with something like insomia which is a genuine diagnosis unless the conversation seems to indicate there is insomnia as a result of poor sleep hygiene. These are self-reported diagnoses so it is possible that the patient misunderstood the diagnosis they received from a clinician and reported 'poor sleep hygiene' as opposed to insomnia. Use the conversation context to make the best judgement.
4. Watch for typos in the diagnosis such as 'golfer elvow' which is really 'Golfer's Elbow'.
5. A diagnosis may contain modifiers like 'slap labrum tear with shoulder instability' which provide more context on a condition which may or may not further distinguish the diagnosis. In this case, it is important to make distinctions when a diagnosis with context actually makes it a new condition that is distinct from the preexisting condition or the resulting condition alone. For example, a 'hemiplegic migraine' is a distinct primary headache disorder that includes motor weakness and should not be considered the exact same diagnosis as a standard 'migraine with aura', even though both fall under the broader category of migraines. However, 'vestibular migraine' and 'migraine-associated vertigo' are the same and both distinct from a standard 'migraine'. Similarly, a patient may have reported a diagnosis with reduced detail like 'arthritis' while the candidate is more detailed like 'Cervical Spondylosis (Arthritis of the Neck)'. If the conversation clearly indicates context that the arthritis is localized to the neck, then this should be treated as a match.
6. A diagnosis may be reported as 'suspected X' where X is the diagnosis. in this case, use the conversation as context to decide whether to match with X if it is present as a candidate. In most cases, this should be treated as a match. Similarly, if a diagnosis contains patient doubt like 'Doctor is telling me I have X' and X is a candidate, threat this as a match as the actual doctor's opinion is more valid over the patient's own doubt.
7. A diagnosis may be unspecific like 'unspecified viral or bacterial infection' in which case it is maybe partially similar to a specific viral or bacterial infection but NOT a perfect match. Please note cases like this.
8. A diagnosis may be a parent condition like 'URI' while the candidates are all children of that like 'influenza'. This should only be made a match if there is clear context within the conversation indicating which specific URI the patient has. If there is no clear indication, do not match a parent condition with a child of that condition without context.
Please format the response in XML strictly using the following order and tags:
<reasoning> - Provide the explanation for why this is a match or why a match was not found. Writing this first helps you reason through the comparison. <position> - Provide the integer index of the matching diagnosis in the differential diagnosis list, starting at 1. If the diagnosis is NOT found in the differential, output -1. <ambiguity_reasoning> - Include any additional notes about ambiguity here. Leave blank if not applicable. <ambiguity> - Output a score from 1 to 5 for how ambiguous the match is.
An example of a clear match might look like this: <reasoning>The self-reported diagnosis of Osteoarthritis perfectly matches the abbreviation 'OA' found at the first position in the differential list.</reasoning> <position>1</position> <ambiguity_reasoning>There is little ambiguity because OA is a standard abbreviation for Osteoarthritis and there are no other conditions which meet this</ambiguity_reasoning> <ambiguity>1</ambiguity>
An example of an ambiguous match might look like this: <reasoning>The reported diagnosis of common cold aligns with the differential candidate at position 4 'unspecified viral or bacterial infection', but the level of granularity between these two diagnoses differs slightly.</reasoning> <position>4</position> <ambiguity_reasoning>While the differential diagnosis does not explicitly provide common cold, it is a viral infection. The fact that the candidate in the differential diagnosis specifically states that the viral infection is 'unspecified' makes this not an exact match. Additionally, the candidate also includes the potential for a bacterial infection. This means the differential candidate is partially correct but lacks the exact granularity of the reported diagnosis.</ambiguity_reasoning> <ambiguity>4</ambiguity>
An example of no match might look like this: <reasoning>The reported diagnosis of 'appendicitis' does not conceptually or textually match any of the gastrointestinal conditions listed in the differential.</reasoning> <position>-1</position> <ambiguity_reasoning>There is no matching candidate in the list or any conditions which could be classified as appendicitis</ambiguity_reasoning> <ambiguity>1</ambiguity>
CRITICAL INSTRUCTION: Provide ONLY the raw XML output. Do not wrap your response in Markdown code blocks (e.g., do not use "'xml). Do not include any conversational text before or after the XML.
Here is the conversation with the differential diagnosis redacted as context: {row['redacted_conversation']} Here is the patient's reported diagnosis. This should be treated as the ground true diagnosis they received: {row['diagnosis_from_survey']} Here is the list of candidate diagnoses in the differential diagnosis they received: 1. {row['diagnosis 1']}, 2. {row['diagnosis 2']}, 3. {row['diagnosis 3']}, 4. {row['diagnosis 4']}, 5. {row['diagnosis 5']}

References

L. Aissaoui Ferhi, M. Ben Amar, F. Choubani, and R. Bouallegue. Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach with clinical vignettes and benchmarking. *Frontiers in artificial intelligence*, 7:1397388, 2024.

R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, 2023.

L. Bastarache, J. C. Denny, and D. M. Roden. Phenome-wide association studies. *JAMA*, 327(1):75–76, 2022.

A. Bean, R. E. Payne, G. Parsons, H. R. Kirk, J. Ciro, R. Mosquera-Gómez, S. Hincape, A. Ekanayaka, L. Tarassenko, L. Rocher, et al. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 2025.

A. M. Bean, R. E. Payne, G. Parsons, H. R. Kirk, J. Ciro, R. Mosquera-Gómez, S. Hincapié M, A. S. Ekanayaka, L. Tarassenko, L. Rocher, et al. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, pages 1–7, 2026.

S. Bedi, H. Cui, M. Fuentes, A. Unell, M. Wornow, J. M. Banda, N. Kotecha, T. Keyes, Y. Mai, M. Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

D. Chambers, A. Cantrell, M. Johnson, L. Preston, S. K. Baxter, S. Baxter, A. Booth, and J. Turner. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. *Health and Social Care Delivery Research*, 7(29):1–88, 2019.

B. Costa-Gomes, P. Tolmachev, E. Taysom, V. Sounderajah, H. Richardson, P. Schoenegger, X. Liu, M. M. Nour, S. Spielman, S. F. Way, et al. Public use of a generalist llm chatbot for health queries. *Nature Health*, pages 1–8, 2026.

N. Crisp and L. Chen. Global supply of health professionals. *New England Journal of Medicine*, 370 (10):950–957, 2014.

Gallup. Americans rely mainly on own doctor for medical advice. URL https://news.gallup.com/poll/702164/americans-rely-mainly-own-doctor-medical-advice.aspx. Accessed: April 22, 2026.

S. Gilbert, H. Harvey, T. Melvin, E. Vollebregt, and P. Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398, 2023.

E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969, 2024.

M. L. Graber. The incidence of diagnostic error in medicine. *BMJ quality & safety*, 22(Suppl 2): ii21–ii27, 2013.

J. R. Hampton, M. Harrison, J. R. Mitchell, J. S. Prichard, and C. Seymour. Relative contributions of

history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *Br Med J*, 2(5969):486–489, 1975.

H. Hayat, M. Kudrautsau, E. Makarov, V. Melnichenko, T. Tsykunou, P. Varaksin, M. Pavelle, and A. Z. Oskowitz. Toward the autonomous ai doctor: Quantitative benchmarking of an autonomous agentic ai versus board-certified clinicians in a real world setting. *arXiv preprint arXiv:2507.22902*, 2025.

R. Heumann and S. R. Steinhubl. Associations between online search trends and outpatient visits for common medical symptoms in the united states from 2004 to 2019: Time series ecological study. *JMIR Formative Research*, 9(1):e77274, 2025.

T. Hirosawa, Y. Harada, M. Yokose, T. Sakamoto, R. Kawamura, and T. Shimizu. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *International journal of environmental research and public health*, 20(4):3378, 2023.

X. Jia, Y. Pang, and L. S. Liu. Online health information seeking behavior: a systematic review. In *Healthcare*, volume 9, page 1740. MDPI, 2021.

Z. Kanjee, B. Crowe, and A. Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80, 2023.

Y. Liu, S. Yu, H. Jin, J. Wen, A. Qian, T. Lee, M. Ramsis, G. W. Choi, L. Qin, X. Liu, et al. A multi-agent framework combining large language models with medical flowcharts for self-triage. *Nature Health*, pages 1–10, 2026.

A. K. Manrai, A. Rodman, et al. Performance of a large language model on the reasoning tasks of a physician. *Science*, 388(6792):123–130, 2026. doi: 10.1126/science.adz4433. URL https://www.science.org/doi/10.1126/science.adz4433.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642 (8067):451–457, 2025.

H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

J. W. O'Sullivan, A. Palepu, K. Saab, W.-H. Weng, D. K. Amponsah, E. Cheng, Y. Cheng, E. Chu, Y. Desai, A. Elezaby, et al. A large language model for complex cardiology care. *Nature Medicine*, pages 1–8, 2026.

A. Palepu, V. Dhillon, P. Niravath, W.-H. Weng, P. Prasad, K. Saab, R. Tanno, Y. Cheng, H. Mai, E. Burns, et al. Exploring large language models for specialist-level oncology care. *NEJM AI*, 2(11): AIcs2500025, 2025a.

A. Palepu, V. Liévin, W.-H. Weng, K. Saab, D. Stutz, Y. Cheng, K. Kulkarni, S. S. Mahdavi, J. Barral, D. R. Webster, et al. Towards conversational ai for disease management. *arXiv preprint arXiv:2503.06074*, 2025b.

M. C. Peterson, J. H. Holbrook, D. Von Hales, N. Smith, and L. Staker. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *Western Journal of Medicine*, 156(2):163, 1992.

E. Riboli-Sasco, A. El-Osta, A. Alaa, I. Webber, M. Karki, M. L. El Asmar, K. Purohit, A. Painter, and

B. Hayhoe. Triage and diagnostic accuracy of online symptom checkers: systematic review. *Journal of Medical Internet Research*, 25:e43803, 2023.

M. Roshan and A. Rao. A study on relative contributions of the history, physical examination and investigations in making medical diagnosis. *The Journal of the Association of Physicians of India*, 48 (8):771–775, 2000.

K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

K. Saab, J. Freyberg, C. Park, T. Strother, Y. Cheng, W.-H. Weng, D. G. Barrett, D. Stutz, N. Tomasev, A. Palepu, et al. Advancing conversational diagnostic ai with multimodal reasoning. *arXiv preprint arXiv:2505.04653*, 2025.

R. Sayres, Y. Hao, A. Ward, A. Wang, B. Freeman, S. Zhan, D. Ardila, J. Li, I.-C. Lee, A. Iurchenko, et al. Towards better health conversations: The benefits of context-seeking. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2026.

M. L. Schmieding, R. Mörgeli, M. A. Schmieding, M. A. Feufel, and F. Balzer. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. *Journal of medical Internet research*, 23(3):e24475, 2021.

H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra. Evaluation of symptom checkers for self diagnosis and triage: audit study. *bmj*, 351, 2015.

Y. Shahsavar, A. Choudhury, et al. User intentions to use chatgpt for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Human Factors*, 10(1):e47564, 2023.

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144, 2024.

J. H. Shen, S. Carter, R. Dargan, J. Gillotte, K. Handa, J. Hong, S. Huang, K. Jagadish, M. Kearney, B. Levinstein, R. Linthicum, M. McCain, T. Millar, M. Julapalli, S. Price, M. Stern, D. Saunders, A. Tamkin, A. Vallone, J. Clark, S. Pollack, J. Eaton, D. Ganguli, and E. Durmus. How people ask Claude for personal guidance, apr 2026. URL https://www.anthropic.com/research/claude-personal-guidance. Accessed: 2026-05-02.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

J. Sumner, Y. Wang, S. Y. Tan, E. H. H. Chew, and A. Wenjun Yip. Perspectives and experiences with large language models in health care: Survey study. *Journal of Medical Internet Research*, 27:e67383, 2025.

T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.

E. Vedadi, D. Barrett, N. Harris, E. Wulczyn, S. Reddy, R. Ruparel, M. Schaekermann, T. Strother, R. Tanno, Y. Sharma, et al. Towards physician-centered oversight of conversational diagnostic ai. *arXiv preprint arXiv:2507.15743*, 2025.

W. Wallace, C. Chan, S. Chidambaram, L. Hanna, F. M. Iqbal, A. Acharya, P. Normahani, H. Ashrafian, S. R. Markar, V. Sounderajah, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ digital medicine*, 5(1):118, 2022.

878 Y. Wang, K. Hunt, I. Nazareth, N. Freemantle, and I. Petersen. Do men consult less than women? an
879 analysis of routinely collected uk general practice data. *BMJ open*, 3(8):e003320, 2013.

880 A. N. Winn, M. Somai, N. Fergestrom, and B. H. Crotty. Association of use of online symptom checkers
881 with patients' plans for seeking care. *JAMA network open*, 2(12):e1918561, 2019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nreditorialpolicychecklistSymptomAI.pdf
- nrreportingsummarycompletedSymptomAI.pdf
- nrsoftwarepolicycompletedSymptomAI.pdf