medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A Novel Playbook for Pragmatic Trial Operations to Monitor and Evaluate Ambient Artificial Intelligence in Clinical Practice

Majid Afshar MD, MSCR[1,2]; Felice Resnik, PhD[1]; Mary Ryan Baumann, PhD[1,3,4]; Josie Hintzke, MS[1]; Anne Gravel Sullivan, PhD[1] Tina Shah, MD, MPH[5]; Anthony Stordalen[5]; Michael Oberst, PhD[5]; Jason Dambach, MD[2,8]; Leigh Ann Mrotek, PhD[1]; Mariah Quinn, MD, MPH[2]; Kirsten Abramson, MD[2]; Peter Kleinschmidt, MD[2]; Tom Brazelton, MD, MPH[6]; Heidi Twedt, MD[2,8]; David Kunstman, MD[7,8]; John Long, MS[7]; Brian Patterson, MD, MPH[8,9]; Frank Liao, PhD[8,9]; Stacy Rasmussen[8], Elizabeth Burnside, MD, MS[1]; Cherodeep Goswami[8]; Joel Gordon, MD[7,8]

[1]Institute for Clinical and Translational Research, School of Medicine and Public Health, University of Wisconsin, Madison, WI

[2]Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI

[3]Department of Population Health Sciences, School of Medicine and Public Health, University of Wisconsin, Madison, WI

[4]Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin, Madison, WI

[5]Abridge AI. Inc, Philadelphia, PA

[6]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin, Madison, WI

[7]Department of Family and Community Health, School of Medicine and Public Health, University of Wisconsin, Madison, WI

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[8]University of Wisconsin Hospitals and Clinics, Madison, WI

[9]BerbeeWalsh Department of Emergency Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI

SOFTWARE VERSION: General Availability release as of October 10, 2024

FUNDING: This work was supported by funding from the University of Wisconsin Hospital and Clinics and the National Institute of Health Clinical and Translational Science Award (NIH/NCATS UL1TR002737). No funding was provided by the AI software company and all licenses of the software were procured as a vendor software as a service (SaaS) Agreement between UW Health and Abridge AI, Inc 2024©.

CORRESPONDING AUTHORS:

Majid Afshar, MD, MSCR

Director, ICTR Learning Health System

Associate Professor of Medicine

600 Highland Ave

UW Health University Hospital

Madison, WI 53792

Joel Gordon, MD

UW Health Chief Medical Information Officer

Associate Professor of Family Medicine and Community Health

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

600 Highland Ave

UW Health University Hospital

Madison, WI 53792

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT

Background: Ambient artificial intelligence offers promise for improving documentation efficiency and reducing provider burden through clinical note generation. However, challenges persist in workflow integration, compliance, and widespread adoption. This study leveraged a Learning Health System (LHS) framework to align research and operations using a hybrid effectiveness-implementation protocol, embedded as pragmatic trial operations within the electronic health record (EHR).

Methods: An alpha phase was conducted to pilot technical integration, refine workflows, and determine sample size in planning for a beta phase designed as a pragmatic randomized controlled trial with the Stanford Professional Fulfillment Index (PFI) as primary outcome. During alpha, bi-directional governance was established between IS operations and LHS team with multidisciplinary workgroups for analytics, technical, documentation, and user experience. Ambient AI was embedded into the EHR using Fast Healthcare Interoperability Resources (FHIR), with real-time data dashboards tracking utilization and documentation accuracy for operations and research. Performance metrics were monitored serially using a difference-in-differences (DiD) analysis to detect drift caused by software workflow changes.

Results: The alpha phase, designed as Type 1 Hybrid, informed a 24-week beta phase stepped-wedge trial with 90% power to detect changes in PFI. Across the alpha phase, the weighted median of average provider Ambient AI utilization was 65.4% following Plan-Do-Study-Act cycles addressing organizational feasibility and task-dependent adoption. Diagnosis code accuracy dropped from 79% to 35% ($p < 0.01$) during alpha but recovered with a new note template and provider training. DiD did not detect significant drifts in work outside of work or

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

time in notes two weeks before and after the new note template. Beta phase enrollment achieved its targeted 66 providers across eight specialties, initiating on schedule.

Conclusions and Relevance: We provide a novel playbook for integrating Generative AI platforms in healthcare, combining pragmatic trial operations, human-centered design, and real-time monitoring to advance evidence-based implementation.

ClinicalTrials.gov ID: NCT06517082

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

INTRODUCTION

The rapid commercialization of generative artificial intelligence (GenAI) tools has outpaced the development of research methods and regulatory oversight to evaluate comprehensive benefits and consequences, creating a critical knowledge gap.[1] This challenge is exacerbated by the fact that health system operations and research organizations often function in silos, creating barriers to evidence-informed implementation of new care pathways.[2] Addressing these barriers requires aligning operational and research priorities through structured governance and workflow-aligned evaluation frameworks. Prior studies show that implementing new tools like GenAI without tailoring for human factors and organizational contexts often results in underutilization and operational inefficiencies.[1,3] A Learning Health System (LHS) framework is well-suited to this challenge, providing the structure to bridge the gap between innovation and effective implementation by enabling continuous improvement through iterative, data-driven cycles of evidence generation and use.[4]

We present a pragmatic randomized controlled trial protocol within a LHS framework to evaluate Ambient AI, a GenAI tool designed to assist with clinical documentation. Clinical documentation in electronic health records (EHR) remains a significant contributor to provider burnout, frequently requiring after-hours "pajama time" to complete.[5,6] Ambient AI aims to reduce this burden, but successful implementation requires achieving workflow integration, ensuring data privacy compliance, and maintaining documentation accuracy. Prior evaluations of Ambient AI have been limited by observational designs, with inconsistent results and limited system interoperability.[7–9] To address these limitations, we designed a pragmatic type 1 hybrid randomized controlled trial within a LHS framework, embedding the Ambient AI system in the

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

EHR. The protocol incorporated governance structures,[10] iterative Plan-Do-Study-Act cycles,[11] and frameworks such as the Systems Engineering Initiative for Patient Safety to ensure scalability, organizational alignment, and human factors considerations.[10,12,13]

Informed by software development principles, the project charter employed alpha and beta testing phases to assess system safety, usability, and clinical impact. The alpha phase served as a controlled environment to refine workflows, address compliance requirements, and establish monitoring processes. Building on these insights, we outline the beta phase clinical trial protocol, structured as a Pragmatic Trial Operations (PTOps) playbook. This novel approach aligned research and operational priorities, integrating activities to create a scalable model for health system-wide deployment. The PTOps playbook advances governance, evidence-based evaluation, and a highly integrated adoption of Ambient AI in clinical practice.

METHODS

Pragmatic Trial Operations (PTOps) Playbook:

The Information Systems (IS) operations project charter and the pragmatic trial protocol were developed using LHS best practices. The PTOps playbook consisted of five core components: governance, user experience, technical, documentation sustainment (i.e., coding compliance), and analytics. The alpha phase aligned with a type 1 hybrid design, while the beta phase implemented a multisite, closed-cohort stepped-wedge pragmatic randomized clinical trial (PRCT) to evaluate Ambient AI's impact on provider well-being. The protocol followed SPIRIT-AI guidelines,[14] and the checklist and full protocol are provided in **Supplemental**.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Setting and Environment

Ambient AI (©Abridge AI, Inc) was integrated into the EHR (©Epic Systems, Inc) under a Software as a Service (SaaS) agreement at the University of Wisconsin Hospitals and Clinics (UW Health). Ambient AI use was part of a quality improvement initiative, with patients providing informed consent during routine care, while providers participating in the PRCT consented to the research component. Since Ambient AI licenses were purchased for clinical use regardless of the trial, the study qualified for expedited review under 45 CFR 46.110 by the UW Institutional Review Board (UW IRB #2024-1028), posing no more than minimal risk. A stepped-wedge design was selected due to the impracticality of rolling out the intervention to all providers simultaneously. This staggered implementation ensured all licenses were eventually deployed, aligning with the operational goals of the quality improvement initiative (**Figure 1B)**. The design also allowed for random concurrent exposure to isolate Ambient AI's effects from secular trends while maintaining guaranteed access for all providers as part of the clinical operations strategy.[15,16]

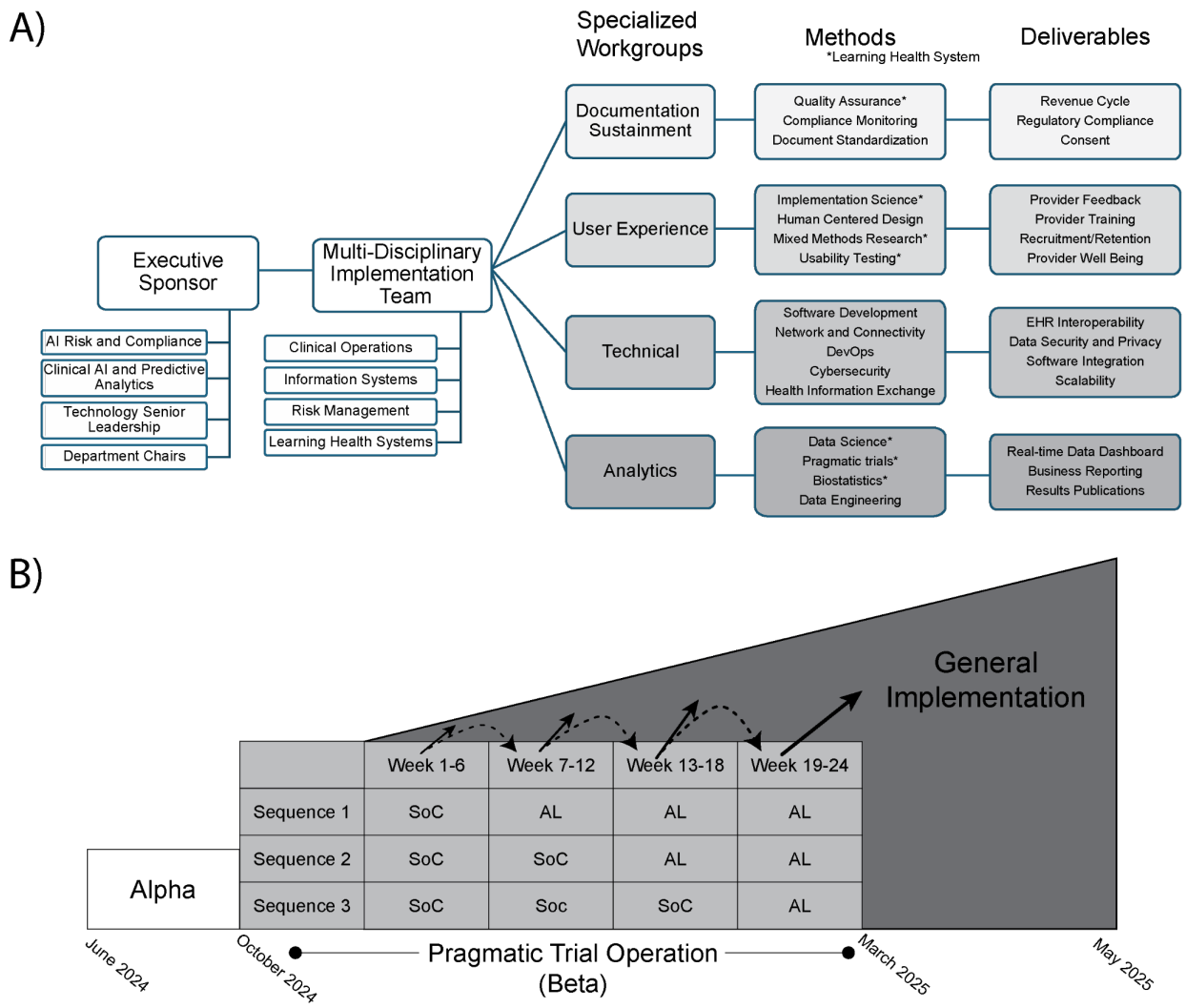
PTOps Playbook Part A: Shared Governance

An architectural framework[17] represented key dimensions (goals, technical, social, ethics, and scientific) to support our LHS ecosystem in which selected projects are implementation-focused, stakeholder engaged, and ready for dissemination. For Ambient AI, a bi-directional governance structure integrated executive sponsors with a Multi-Disciplinary Implementation Team comprising leaders from clinical operations, IS, health information management, AI risk and compliance, and LHS programs. Specialized workgroups addressed analytics, technical, documentation sustainment, and user experience (**Figure 1A**). The governance committee met

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

biweekly to resolve protocol deviations, assess risks and benefits, allocate resources, monitor safety, and address unanticipated changes in the implementation timeline.

Figure 1. Governance across Alpha and Beta Phases with Pragmatic Trial Operations



AL = Ambient AI Learning; SoC = Standard of Care, also known as usual care. In pragmatic trial operations, if the implementation proves to be ineffective from the clinical trial then de-implementation and repeat of the LHS framework can be employed before moving to general implementation.

PTOps Playbook Part B: User Experience

Alpha Phase: User experience was informed by the Exploration, Preparation, Implementation, and Sustainment (EPIS)[18] and modified Systems Engineering Initiative for Patient Safety

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(SEIPS)[13] frameworks. The EPIS framework highlighted the interplay of contextual factors, including outer context (organizational readiness), inner context (provider characteristics), innovation (Ambient AI), and bridging factors (communication and coordination). The SEIPS model, with its People, Environment, Tools, and Tasks (PETT) Scan,[19] offered a practical approach for mapping work processes and identifying potential barriers and facilitators to mitigate cognitive load (**Supplemental**).

Beta Phase: Department leaders distributed surveys to identify eligible providers to facilitate efficient recruitment and enrollment for the PRCT. Training materials and workflows were iteratively refined based on provider feedback from the rapid Plan-Do-Study-Act cycles during the alpha phase.

PTOps Playbook Part C: Technical Strategies

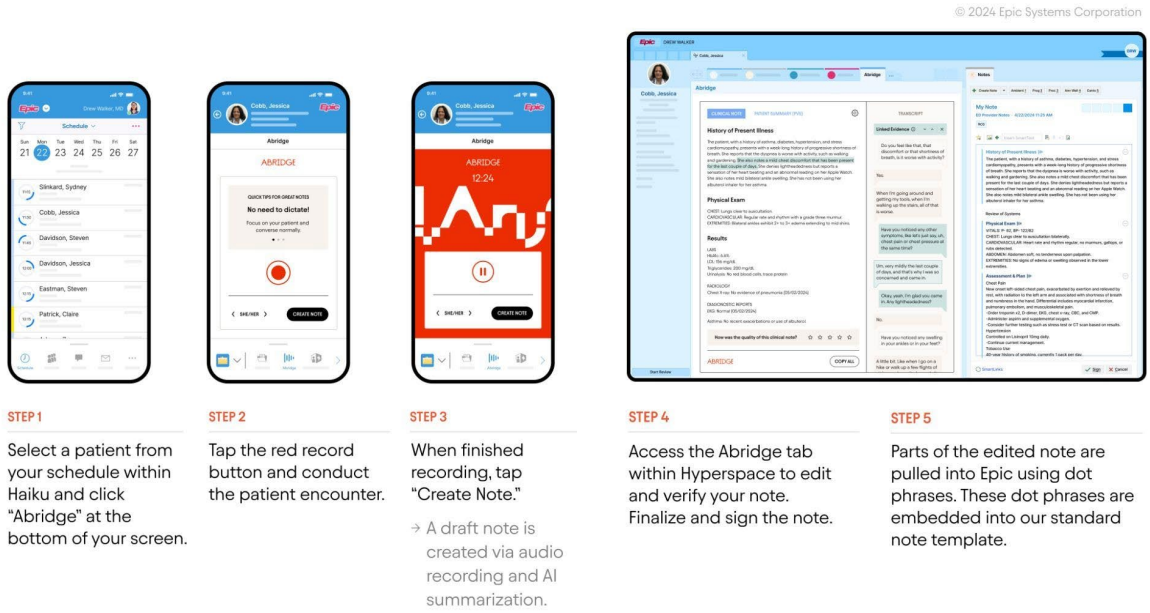
Alpha Phase: EHR-embedding of Ambient AI utilized Epic's Private Application Programming Interface (APIs) and Fast Healthcare Interoperability Resources (FHIR) R4 APIs to transfer AI-generated notes and session metadata into the EHR. Providers obtained verbal consent from patients and used the Epic Haiku mobile application to start the recording (**Figure 2**). The technical workgroup coordinated access, and deployment aligned with the PRCT randomization schedule. Daily reviews addressed consent compliance and system performance.

Beta Phase: PRCT inclusion criteria were providers seeing at least 20 patients per week, who had completed the required training, and who used Epic Haiku. Providers with planned leave, unsupported mobile devices, or reluctance to discontinue the use of existing medical scribes were

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

excluded. The study statistician generated randomization allocation lists for all waves prior to the initiation of the PRCT. To align with the PRCT, operational staff coordinated with the research team to contact providers, schedule training, and oversee technical onboarding.

Figure 2. Abridge Ambient software integration with Epic electronic health record system.



PTOps Playbook Part D: Documentation Sustainment

Alpha Phase: Certified coders and informaticists monitored diagnostic codes and manually reviewed signed provider documentation for adherence to coding standards and regulatory requirements, including diagnoses mentioned in the AI-generated note that were not in the structured data, visit diagnosis list, or lacked specificity. Due to unanticipated decreases in diagnostic coding accuracy, two strategies were employed: (1) development of a new reporting template; and (2) automated extraction. The latter method used billing codes in the EHR to automatically extract International Classification of Diseases (ICD)-10 codes using a large language model (LLM) developed by UW. The LLM was designed with four prompt engineering

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

strategies: minimizing perplexity,[20] in-context examples, chain-of-thought, and self-consistency.[21] The prompt is shown in **Supplemental**.[22]

Beta Phase: The LLM system was validated using OpenAI's Generative Pre-Trained Transformer 4 omni (GPT-4o), operating within a secure, HIPAA-compliant Azure cloud environment. EHR data sent to GPT-4o adhered to HIPAA regulations, ensuring patient confidentiality. To evaluate the LLM's accuracy, the Jaccard index was employed to measure the similarity between ICD-10 codes extracted from AI-generated notes by the LLM and those manually finalized by certified hospital coders. A Jaccard score of 1 indicated perfect alignment between LLM-extracted codes and coder-reviewed codes. Weekly monitoring by the LLM system flagged cases for review by the coders.

PTOps Playbook Part E: Outcomes-oriented Analytics

Alpha Phase: Early collaboration with the Chief Wellness Officer facilitated the adoption of the Professional Fulfillment Index (PFI)[23], a validated tool routinely used at UW Health to measure provider well-being. Historical PFI survey data from 1,091 providers at UW Health in 2023 informed the PRCT's power and sample size calculations. The LHS team estimated a within-period intracluster correlation coefficient (ICC) of 0.032 for overall PFI scores and 0.034 for its burnout subcomponent. Provider autocorrelation was estimated at 0.65, and the clinically meaningful effect size was set at Cohen's d of 0.44. Based on these parameters, the LHS team provided precise recommendations on the minimum sample size. Randomization of 66 providers was planned into three waves to achieve 90% power to detect a meaningful difference in PFI.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A real-time data dashboard (©1993-2024 QlikTech International AB) integrated data from the EHR and Abridge software to monitor process and efficiency measures of provider activity (**Figure 3**).[24] Dashboard metrics included provider feedback, utilization rates, patient consent compliance, and documentation efficiency. This last metric was quantified by work outside of work (WoW) and time spent on notes (both normalized to the 8-hour workday) as well as closures before next encounter by end of day, and patient follow ups.[24] Since licenses were issued at the provider level, the weighted medians of average daily provider-level metrics were measured to capture central tendencies and variability.[25] Weights were created by a provider's relative number of in-clinic days compared to the total number of days. REDCap[26] surveys were designed to collect the PFI and provider-reported outcomes (**Supplemental Clinical Trial Protocol**).

Figure 3. Components of real-time data dashboard in operations

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The data displayed are alpha users only and not part of the clinical trial beta phase. These are for demonstration purposes only.

Beta Phase: The Beta phase was designed as a 24-week, individually randomized, stepped-wedge trial (**Figure 1B**). Providers were randomized 1:1:1 into three waves using stratified permuted-block randomization, ensuring balance across clinical specialties. Each wave transitions from practice-as-usual to the intervention at 6-week intervals to provide sufficient time for the Ambient AI's effect to manifest with intent-to-treat principles. The statistical analysis will include random effects for individual providers and time interactions, along with fixed effects for time periods to adjust for secular trends.

The novelty of Ambient AI technology and absence of predefined adverse event categories for software changes necessitated innovative monitoring strategies. A difference-in-differences (DiD) analysis was implemented with a rolling two-week time window to identify unexpected drifts in performance. The primary metrics monitored were utilization, WoW, and time in notes. For each metric, trends for the AI-generated encounters over a two-week period were compared to trends for the non-AI encounters. Analyses employed linear mixed-effects models with random intercepts for individual providers and fixed effects for time, accounting for autocorrelation across two-week intervals. Statistical significance was defined as a Bonferroni-corrected p-value ≤ 0.05, adjusted for the total number of tests performed. Significant drifts triggered root cause analyses and discussions with operational leaders. The top 2% of WoW and time in notes observations were excluded from analysis as artifacts that were deemed clinically implausible.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

RESULTS

The governance structure successfully aligned operational and research priorities, integrating the Ambient AI trial within the health system processes while positioning ambient listening for widespread adoption if proven effective. Endorsement was received from leadership and AI oversight committees (**Figure 1A)**.

The alpha phase started on June 24, 2024 until the initiation of the beta phase clinical trial on October 10, 2024. A total of 20 providers with 8,527 clinic encounters were evaluated during the alpha phase. The providers were distributed across five specialties, 12 clinic locations, and 50% were female. A control chart was developed using alpha phase user data to establish the utilization threshold for Ambient AI with a lower control limit of three standard deviations from the mean set at 48%, guiding license allocation and tracking fidelity.

During the alpha phase, ten providers (50%) participated in interviews with the LHS team, sharing both positive and constructive feedback about their experiences with Ambient AI. Positive comments included statements such as, “I feel like I was walking before, and now this is like a bullet train, and a scribe would be something like a stagecoach”. However, challenges were also highlighted, such as, “I have had some patients decline citing privacy concerns even after discussing how it works and that it’s compliant.” Key themes emerged from the interviews: (1) organizational characteristics, including clinic-level physical environments and team dynamics, influenced Ambient AI utilization; (2) provider and patient readiness, as well as documentation preferences, contributed to variability in adoption; and (3) the suitability of Ambient AI platforms to offload providers depended on service settings and clinical tasks. To

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

address these barriers, interviewers used the SEIPS PETT Scan framework to map the identified facilitators and barriers across different parts of the health system that resulted in recommendations to improve fidelity (**Table 1**). Rapid Plan-Do-Study-Act cycles were employed to address issues such as inconsistent use of note templates and gaps in obtaining patient consent. These iterative modifications included updating training materials, enhancing technical support, and refining workflows. Across the alpha phase, the weighted median of provider-averaged daily Ambient AI utilization was 65.4% (IQR 50.6% - 84.0%).

Table 1. Key Themes mapped to SEIPS PETT Scan with Examples of Barriers and Facilitator to Inform Implementation**(B= Barriers F = Facilitators)**

Theme	People	Environments	Tools	Tasks
1. Organizational characteristics	-Time pressure$_{B}$ -Documentation pressure/stress$_{BF}$	-Diverse physical settings$_{B}$ -Centralized implementation team$_{F}$ -Standardized workflows$_{F}$	-Internet/server connectivity issues$_{B}$	-Technical assistance$_{BF}$
2. Provider and patient readiness (individual characteristics)	-Understanding, willingness, and learning curve$_{B}$ -Provider documentation preferences$_{B}$ -Excited about new technology$_{F}$	-Preferences for specific clinics$_{B}$ -App burden$_{B}$	-Quirk tolerance for tool$_{BF}$	-More connection and eye-contact with patients$_{F}$
3. Type of service setting and clinical task needs	-Acceptability/usability of tool and generated note$_{BF}$	-Templates for state specific requirements$_{B}$	-Templates for specialties & types of visits$_{B}$ -Technology quirks/glitches$_{B}$ -Integration with EHR$_{F}$	-Not able to do certain tasks (pending orders, pre-visit summaries)$_{B}$ -Compliance issues$_{B}$
Suggested Implementation Strategies and Adaptations (EPIS phase) • Audit and provide feedback (IS) • Booster training sessions (PIS) • Centralized and local technical assistance (PIS) • Clear guidance (PIS) • Communicate data to external stakeholders, providers, and patients to demonstrate ongoing benefit (IS) • Information for patients (PIS) • Positive messaging and testimonials about value (EPIS) • Stage implementation scale-up (PI) • Tools for quality monitoring (PIS) • Training materials and FAQ sheets for providers (PIS) • Training and reminders for compliance (IS)				

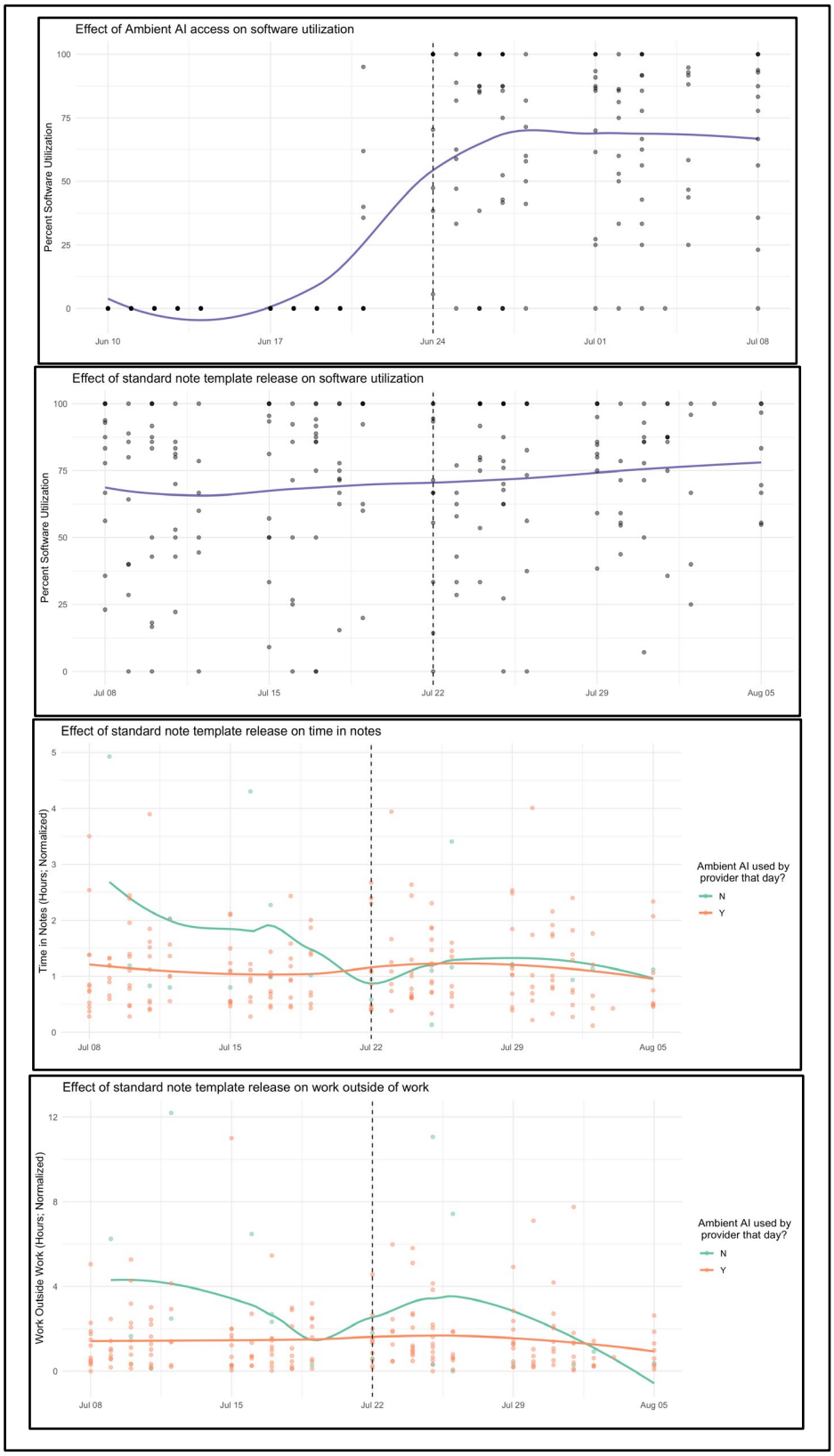
Between June 24, 2024 and July 8, 2024, Ambient AI utilization increased by an average of 57.2% ($p<0.01$), which aligns with the release of the licenses to alpha providers.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Manual review of 797 visit encounters across 17 providers identified discrepancies in diagnosis codes between provider ICD-10 coding and Ambient AI-generated coding. With the transition from diagnosis-linked manual charting to AI-generated prose for the assessment and plan, accuracy of ICD-10 diagnosis codes decreased from 79% pre-Ambient AI to 35% post-Ambient AI ($p<0.01$) with example notes shown in **Supplemental**. Following the implementation of a new note template on July 20th, designed to align documentation with institutional coding practices, most notably by placing the visit diagnosis into the attestation section, documentation accuracy improved and returned to pre-Ambient AI levels. As part of a novel monitoring strategy to detect significant drifts in provider metrics, no substantial drift was observed across a two-week rolling window. Using difference-in-differences analysis to compare encounters with AI-generated notes to usual care controls, results showed an average decrease in Ambient AI utilization of 6.84 percentage points (SD=5.95), an average increase in work outside of work by 0.33 hours (SD=0.78), and an average increase in time spent in notes by 0.43 hours (SD=0.25) (**Figure 4**).

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4. Differences-in-differences across 2-week window for Ambient AI initiation and new standard note implementation



medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

During the alpha phase, coders logged a total of 598.33 hours manually reviewing diagnosis documentation accuracy. Recognizing the unsustainability of manual review for the larger cohort of providers in the beta phase, an automated coding review system leveraging a LLM assistant was developed. Validation of 200 clinic notes with ICD-10 diagnosis codes manually reviewed by certified hospital coders demonstrated that the LLM assistant achieved a mean Jaccard score of 0.54 (95% CI: 0.50–0.58). Following validation, coders conducted iterative reviews to establish an operational Jaccard score threshold between provider-assigned and LLM-generated ICD-10 codes. This threshold served as an initial screen to flag clinic notes for follow-up, effectively prioritizing high-risk cases and reducing the manual review workload.

A pre-recruitment process minimized disruption to clinical workflows by enabling providers to self-identify interest in Ambient AI. Eligibility was based on trial criteria. Following IRB approval in August 2024, a waitlist streamlined recruitment, allowing the clinical trial to proceed without delays. By October 2024, recruitment of the planned 66 providers across eight specialties was completed, spanning sites in Wisconsin and Illinois. The 24-week trial began on schedule. Post-implementation, the data dashboard transitioned to operational ownership, ensuring operational stewardship.

DISCUSSION

This study demonstrates how aligning research and operational priorities within a LHS framework can accelerate the implementation and evaluation of rapidly iterating GenAI technology. A key to success was the governance structure, which supported iterative

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

improvement cycles focused on translating performance to data, data to knowledge, and knowledge to actionable outcomes.[4] These cycles established an infrastructure that balanced the needs of clinical operations efficiency with research rigor, fostering a scalable and sustainable model for evaluating a dynamic GenAI platform.

The alpha phase identified barriers through rapid PDSA cycles, addressing issues such as workflow variability and inconsistent documentation. Unlike previous studies on Ambient AI,[7,8] we built and assessed an Ambient AI platform that was fully embedded within the EHR, allowing for the evaluation of pragmatic workflow issues. Tailored training materials, workflow adjustments, lower control limits for software utilization, and audit mechanisms were developed to mitigate these challenges, achieving high fidelity by alpha phase conclusion. These efforts set a strong foundation for the beta phase and aligned with three key characteristics of socio-technical infrastructure:[27] (1) engaging multi-stakeholder learning communities; (2) rigorously exploring uncertainty during the alpha phase; and (3) fostering sustainability through iterative PDSA cycles. Tailored resources, including FAQs, EHR note templates, training videos, and a dedicated help desk ensured alignment between research objectives and operational processes.

Informatics innovations included automating ICD code review with LLMs and integrating real-time data feeds from the EHR into an operational dashboard. These tools fused research with operations, reduced reliance on manual data collection and supported a scalable evaluation model. The automated LLM-based ICD code review served as a first-pass filter, reducing the burden on coders tasked with manually reviewing all notes. Another key innovation was adapting trial designs to the dynamic nature of GenAI systems. Unlike static interventions in

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

traditional trials, GenAI platforms continue to have updates that may lead to significant drifts in performance. The lack of industry predefined adverse event categories for what constitutes major or minor software changes necessitated our own monitoring strategy to detect unexpected drifts during deployment. The DiD analysis, conducted in short, 2-week intervals, was derived to help fill potential gaps in routine monitoring strategies to address the dynamic nature of AI adoption.

Embedding research rigor into operational decision-making was another strength of this study. Validated metrics like the PFI, aligned with Chief Wellness Officer goals, ensured outcomes were appropriately powered to detect meaningful impact for the clinical trial. This approach avoided reliance on loosely defined metrics, such as time savings, and fostered evidence-based decision-making for the executive sponsors to determine number of initial licenses and duration of evaluation period. Operationalization of the PRCT was strengthened by a multidisciplinary governance structure that managed technical integration, documentation compliance, and analytics. Centralized communication and shared responsibilities facilitated project management, integrating research into clinical workflows with minimal burden on the frontlines.

The dual-purpose data dashboard, developed during the alpha phase and transitioned to operational ownership, exemplified the integration of research and operations. It tracked vital metrics such as Ambient AI utilization, documentation efficiency, drift, and consent compliance while simultaneously supporting research analyses. Functioning as an analog to a Data Safety and Monitoring Board, the dashboard addressed both the immediate needs of the trial and the long-term goals of monitoring and scalability. Our implementation of governance provided oversight from institutional AI committees.[10] Additionally, an operational-research Git

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

repository was established to enhance reproducibility and facilitate compliance with National Institutes of Health data-sharing requirements [*GitLab repository available upon acceptance*]. The governance model is an important component of the playbook for navigating the safety concerns with GenAI implementation.

This study highlights the potential of LHS frameworks to embed PRCTs into an operational timeline, prepare for widespread adoption, and evaluate outcomes with research-grade analytics. By redesigning trial stages of planning, recruitment, outcome ascertainment, and dissemination, this approach accelerates the evaluation of healthcare technologies while maintaining alignment with clinical practice. Importantly, we demonstrated that rigorous, research-grade evaluation can coexist with operational efficiency, addressing concerns from operational leaders about potential delays. As GenAI technologies continue to evolve, this study provides a playbook for bridging the gap between innovation and real-world application, supporting the safe, effective, and sustainable adoption of AI in clinical care.

ACKNOWLEDGEMENTS: Tom Wise (Business Relation Management), Christine Cunningham (Health Information Management), Troy Lepein (Risk and Compliance/Business Integrity), Karen Nachman, Luke Rislove and Sarah Fardy (Enterprise Analytics), Rachelle Buol and Tori L McKinley (Clinical Documentation Integrity), Nicole Riechers (IS Project Management Office), Lisa Wilson and Monica Esquibel (Institutional Review Board), Betsy Nugent and Nasia Safdar (Clinical Trials Institute), Graham Wills (Applied Data Science)

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. *Lancet Lond Engl*. 2022;399(10325):620. doi:10.1016/S0140-6736(22)00235-5
2. Angus DC, Huang AJ, Lewis RJ, et al. The Integration of Clinical Trials With the Practice of Medicine: Repairing a House Divided. *JAMA*. 2024;332(2):153. doi:10.1001/jama.2024.4088
3. Carayon P, Hoonakker P, Hundt AS, et al. Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. *BMJ Qual Saf*. 2020;29(4):329-340. doi:10.1136/bmjqs-2019-009857
4. Smith MA, Adelaine S, Bednarz L, Patterson BW, Pothof J, Liao F. Predictive Solutions in Learning Health Systems: The Critical Need to Systematize Implementation of Prediction to Action to Intervention. *NEJM Catal*. 2021;2(5). doi:10.1056/CAT.20.0650
5. Harry E, Sinsky C, Dyrbye LN, et al. Physician Task Load and the Risk of Burnout Among US Physicians in a National Survey. *Jt Comm J Qual Patient Saf*. 2021;47(2):76-85. doi:10.1016/j.jcjq.2020.09.011
6. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann Fam Med*. 2017;15(5):419-426. doi:10.1370/afm.2121
7. Liu TL, Hetherington TC, Dharod A, et al. Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study. *NEJM AI*. Published online November 22, 2024. doi:10.1056/AIoa2400659
8. Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catal*. 2024;5(3). doi:10.1056/CAT.23.0404
9. Shah SJ, Devon-Sand A, Ma SP, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J Am Med Inform Assoc*.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Published online December 5, 2024:ocae295. doi:10.1093/jamia/ocae295

10. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes. *Front Digit Health*. 2022;4:931439. doi:10.3389/fdgth.2022.931439

11. Plan-Do-Study-Act (PDSA) Directions and Examples. Accessed September 24, 2024. https://www.ahrq.gov/health-literacy/improve/precautions/tool2b.html

12. Carayon P, Wooldridge A, Hoonakker P, Hundt AS, Kelly MM. SEIPS 3.0: Human-centered design of the patient journey for patient safety. *Appl Ergon*. 2020;84:103033. doi:10.1016/j.apergo.2019.103033

13. Carayon P, Wooldridge A, Hoonakker P, Hundt AS, Kelly MM. SEIPS 3.0: Human-Centered Design of the Patient Journey for Patient Safety. *Appl Ergon*. 2020;84:103033. doi:10.1016/j.apergo.2019.103033

14. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7

15. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28(2):182-191. doi:10.1016/j.cct.2006.05.007

16. Federico CA, Heagerty PJ, Lantos J, et al. Ethical and epistemic issues in the design and conduct of pragmatic stepped-wedge cluster randomized clinical trials. *Contemp Clin Trials*. 2022;115:106703. doi:10.1016/j.cct.2022.106703

17. Lessard L, Michalowski W, Fung-Kee-Fung M, Jones L, Grudniewicz A. Architectural frameworks: defining the structures for implementing learning health systems. *Implement Sci*. 2017;12(1):78. doi:10.1186/s13012-017-0607-7

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

18. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement Sci IS*. 2015;10:21. doi:10.1186/s13012-015-0209-1

19. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. *BMJ Qual Saf*. 2021;30(11):901-910. doi:10.1136/bmjqs-2020-012538

20. Gonen H, Iyer S, Blevins T, Smith N, Zettlemoyer L. Demystifying Prompts in Language Models via Perplexity Estimation. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics; 2023:10136-10148. doi:10.18653/v1/2023.findings-emnlp.679

21. Wang X, Wei J, Schuurmans D, et al. SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS. Published online 2023.

22. Afshar M, Gao Y, Wills G, et al. Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record. *JAMIA Open*. 2024;7(3):ooae080. doi:10.1093/jamiaopen/ooae080

23. Trockel M, Bohman B, Lesure E, et al. A Brief Instrument to Assess Both Burnout and Professional Fulfillment in Physicians: Reliability and Validity, Including Correlation with Self-Reported Medical Errors, in a Sample of Resident and Practicing Physicians. *Acad Psychiatry*. 2018;42(1):11-24. doi:10.1007/s40596-017-0849-3

24. Sinsky CA, Rule A, Cohen G, et al. Metrics for assessing physician activity using electronic health record log data. *J Am Med Inform Assoc*. 2020;27(4):639-643. doi:10.1093/jamia/ocz223

25. Beliakov G, Bustince H, Fernandez J. The median and its extensions. *Fuzzy Sets Syst*. 2011;175(1):36-47. doi:10.1016/j.fss.2011.01.002

medRxiv preprint doi: https://doi.org/10.1101/2024.12.27.24319685; this version posted December 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010

27. Friedman CP, Lomotan EA, Richardson JE, Ridgeway JL. Socio-technical infrastructure for a learning health system. *Learn Health Syst*. 2024;8(1):e10405. doi:10.1002/lrh2.10405