

A Randomized Controlled Trial of Ambient Artificial Intelligence Scribe Technologies
NCT06792890
2025-01-28

Protocol Paper incl. Statistical Analysis Plan

Title: A Randomized Controlled Trial of Two Ambient Artificial Intelligence Scribe
Technologies to Improve Documentation Efficiency and Reduce Physician Burnout

Aaron T. Chin, MD*^Φ

Department of Pediatrics, Division of Immunology, Allergy and Rheumatology and Health
Department of Information Services and Solutions, University of California, Los Angeles

William Turner, BA^Φ

Division of General Internal Medicine and Health Services Research, David Geffen School
of Medicine, University of California, Los Angeles

Eric M. Cheng, MD, MS

Health Department of Information Services and Solutions, University of California, Los
Angeles

Paul Lukac, MD, MBA, MS

Department of Pediatrics and Office of Health Informatics and Analytics, University of
California, Los Angeles, California, United States

Catherine A. Sarkisian, MD MSHS

Division of General Internal Medicine and Health Services Research, David Geffen School
of Medicine, University of California, Los Angeles and VA Greater Los Angeles Healthcare
System/Geriatrics Research Education and Clinical Center, Los Angeles, California

Ya-Chen Tina Shih, PhD

Department of Radiation Oncology, David Geffen School of Medicine, University of
California, Los Angeles

Sitaram Vangala, MS

Department of Medicine Statistics Core, David Geffen School of Medicine, University of
California, Los Angeles

John N. Mafi, MD, MPH

Division of General Internal Medicine and Health Services Research, David Geffen School
of Medicine, University of California, Los Angeles

*Corresponding Author: atchin@mednet.ucla.edu

^Φ Denotes co-first authors; equal contributors

Abstract

Background: Ambient artificial intelligence (AI) scribe technologies have shown early promise in alleviating documentation burden and physician burnout. However, to our knowledge, these tools have not been rigorously evaluated in a randomized controlled trial (RCT).

Methods: This is a three-arm pragmatic RCT of 238 outpatient physicians at a large academic health system, randomized 1:1:1 to one of two ambient AI scribe tools or a usual-care control group. The two-month study will compare the effects of each tool. We will use covariate-constrained randomization to balance the arms in terms of physician baseline time in notes, survey-measured level of burnout, and clinic days per week.

Study Outcomes: The primary outcome is the change from baseline in physician time spent on documentation. Self-reported physician burnout, task load, professional fulfillment, NET promotor score, and perceived harmful/unintended consequences as well as EHR-based after-hour time, pajama time, and productivity metrics (visits per week, relative value units (RVUs), clinic turnover rates) will be assessed as secondary outcomes. Data will be collected via provider surveys, electronic health record (EHR) utilization data, and provider-level billing data.

Conclusion: To our knowledge, this is the first RCT evaluating two ambient AI scribe vendors. These findings will provide insights into their impact on efficiency, productivity and physician well-being in ambulatory care settings.

Trial Registration: This study is registered on ClinicalTrials.gov under the identifier NCT06792890 (<https://clinicaltrials.gov/study/NCT06792890>)

Keywords: Artificial intelligence, ambient scribe, digital scribe, clinical informatics,
physician burnout

Protocol Version

This is Protocol Version 1, dated November 1, 2024.

Trial Sponsor

Not applicable. This study does not have an external sponsor.

Background

Nearly one in two U.S. physicians experiences some level of burnout, accelerating physician turnover and doubling the risk to patient safety.(1-5) For every hour spent in direct patient care, physicians spend approximately two hours on electronic health record (EHR) documentation, including 1-2 hours per day of after-hours work. This time burden not only reduces face-to-face time with patients but also accelerates physician turnover and contributes to cognitive strain and diminished job satisfaction.(6) Time pressures often lead physicians to expedite documentation by copying prior notes, resulting in 'note bloat' and inaccuracies.(7) To date, this problem remains unresolved.

While interventions such as human scribes and speech recognition software have alleviated some of this burden, both have limitations of cost, availability, and accuracy. (8-11) Leveraging recent rapid advancements in large language models (LLMs), multiple health tech vendors have developed generative artificial intelligence (genAI) -powered solutions to enhance ambient scribes at a lower cost than human scribes. These technologies capture physician-patient conversations to create a transcript, then

summarize the transcript in the form of a clinical note. Early implementations of these models show promise in reducing documentation burden, enhancing patient satisfaction, and improving physician engagement (12-17). However, the implementation of ambient AI scribe technologies raises specific challenges, including variability in performance across different clinical environments, potential equity implications for safety net providers who may not be able to afford these technologies, and uncertainty in their cost-effectiveness.(18) In addition, there are concerns about the use of LLMs use in healthcare, including the loss of nuanced clinical reasoning and hallucinations in AI-generated notes, which could inadvertently increase physician documentation time and burnout. These challenges underscore the importance of rigorous empirical evaluations of ambient AI scribes.(19)

To our knowledge, no studies have conducted a randomized controlled trial (RCT) of ambient AI scribe vendors compared with a usual-care control group. Head-to-head comparison of two ambient AI scribe tools adds to the novelty of this trial. In accordance with the Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence (SPIRIT-AI) guidelines, this study aims to evaluate the effect of two separate ambient AI scribes on reducing physician documentation time, mitigating burnout, and enhancing productivity.(20) We hypothesize that each tool will be associated with significant reductions in time-in-note as well as physician burnout rates.

Methods

Study Aims

The primary aim of this study is to evaluate the impact of two ambient AI scribe technologies on physician change from baseline time spent on EHR documentation, comparing each scribe to a control group. Secondary objectives include assessing the ambient AI scribes' impact on physician metrics such as burnout, physician satisfaction, perceived harmful consequences, and productivity. Additionally, the study team intends to perform an economic evaluation analysis of the tools to guide business decision making.

Participants

Eligible participants for this study included ambulatory care physicians across multiple specialties within the UCLA Health system who held at least one half-day of clinic per week. Trainee providers (e.g., residents, medical students) and allied healthcare professionals (e.g., RNs, PAs) were excluded, along with attendings who work exclusively with trainees and provide no direct patient care. Only English-language encounters are included due to a lack of internal validation for translation capabilities in the vendor technologies. Physicians who typically use a human scribe agreed to forego this assistance while assigned to an intervention or control group. Due to the technical requirements of one vendor, all providers are required to use an iOS portable device.

Participants were recruited based on specific department leadership suggestions and non-specific department-wide emails, resulting in 331 provider invitations. Of these, only 238 met the study criteria. Reasons for exclusion included ineligibility (e.g., did not meet inclusion criteria; $n = 7$), refusal to participate ($n = 21$), status as a key decision-maker (to prevent potential conflicts of interest in outcome assessment; $n = 20$), or failure to

complete the baseline survey (n = 43). Participants who met eligibility requirements and agreed to participate were required to complete a pre-implementation survey prior to randomization.

Study Design and Randomization

Eligible physicians (n = 238) were randomized into one of three groups: either of the two ambient AI scribe technologies (Vendor A or Vendor B) or a control group with no scribe (Figure 1). Randomization followed a 1:1:1 allocation over a two-month period and analysis will follow an intention-to-treat approach. Since both vendor tools are being rolled out from November 4, 2024 to January 3, 2025, secular trends, seasonal variations, and clinic holiday closures are accounted for by comparing effects of Vendor A and Vendor B with results from a contemporaneous control group.

Covariate-constrained randomization was conducted by the trial statistician to balance participants' baseline geometric mean time in notes per note collected over the prior year, as well as burnout scores and clinic days per week obtained from survey data.

Intervention and Implementation

The RCT encompasses an 8-week period from November 4, 2024 to January 3, 2025. Physicians using ambient AI scribe tools will continue their usual clinical documentation processes, supported by the scribe software, which integrates with the EHR and automatically adds the generated text to the note. Training for each product was conducted 1–2 weeks prior to the intervention rollout and was standardized in terms of the time

allotted for each vendor and the format of internally generated support material. This included tipsheets, a recorded webinar, and two 1-hour-long virtual sessions for each vendor to demo the software functionality and workflow. Participants were free to discontinue their assigned intervention at any time. Reasons for discontinuation will be tracked and documented, and all participants will be analyzed on an intention-to-treat basis. To ensure participants adhered to the intervention protocols, comprehensive technical support was provided throughout the study. This included direct post-implementation support (dedicated communication channels, email listservs, and vendor contact), and regular check-ins by the UCLA IT team to help resolve any technical issues.

In compliance with California state law, all physicians must inform patients about the audio recording and obtain their verbal consent. Vendors do not retain any patient data, and recordings are not stored after transcription, following standard data protection and privacy protocols. Data from Epic System's Signal platform were collected starting six months before the 'go-live' date to establish baseline physician characteristics, like time spent writing a note, and for use in future outcomes analyses. At the end of the 2-month intervention period, participants will complete a post-implementation survey, which includes original survey questions and additional items to capture subjective impressions of each tool. Participant dropout is tracked throughout the study period, with reasons for discontinuation documented.

Study Outcomes

Primary Outcome

The primary outcome measure is the change in the time in notes per note in the second month of the trial from six months prior. The second month of the intervention was chosen to allow providers to familiarize themselves with the tools and workflow. This change will be computed on the natural log scale.

Secondary Outcomes

Secondary outcomes include burnout scores from the Mini-Z 2.0, provider task load adapted from the NASA Task Load Index, Professional Fulfillment Index scores, and the Net Promoter Index.(21-23) Additional outcomes include self-reported satisfaction with the documentation method, self-reported effects on note accuracy, patient safety, equity, and other potential unintended consequences, such as hallucinations, harmful, or poor-quality transcriptions. The study team is also recording physician-level billing information via RVU and clinic turnover rate to determine changes in productivity. We will examine change in additional Signal metrics including pajama time per scheduled day, time outside scheduled hours per scheduled day, and number of unscheduled days where time is spent in the system.

Data Collection Methods

The study dataset includes both data already collected via a pre-implementation survey and post-interventional data comprised of Epic data and a follow-up survey. The pre-implementation survey, sent to providers in all three arms through Qualtrics one month prior to implementation, confirmed eligibility and opt-in to the trial. Questions included

baseline assessment of the Mini-Z 2.0, provider task load, and Professional Fulfillment index, as well as provider demographic questions. Post-intervention data collected immediately after implementation will include efficiency metrics provisioned by Epic via its Signal platform, follow-up survey responses, provider-level professional billing information, and encounter-level time data. The follow-up survey will also capture physician-reported effects on note accuracy, patient safety, equity, and other unintended consequences, such as hallucinations or errors in transcription. No patient health information is collected in this study. Table 1 outlines each measured outcome, and the corresponding method used for data collection.

Statistical Analysis

This study protocol follows SPIRIT-AI and CONSORT-AI guidelines for development and reporting. Descriptive statistics will be used to characterize the demographics and baseline information of providers in each arm.

The primary analysis will evaluate the impact of each intervention arm (Vendor A and Vendor B) compared to the control group. Specifically, the two primary comparisons are (1) Vendor A vs. Control and (2) Vendor B vs. Control. The primary outcome, the change in second month's (log-transformed) time in notes per note from the provider's prior six-month baseline, will be analyzed using a linear mixed-effects model. This model will include a study arm effect, a period effect (second vs. first month), and their interaction. Random physician effects will account for multiple timepoints per provider, and linear contrasts will evaluate differences between arms at the second month timepoint. We are

studying the period effect under the assumption that first month represents a learning phase and may not accurately reflect potential efficiency gains, while second month reflects proficiency with the tool. A two-sided 0.025 significance level (2-fold Bonferroni correction) will be used to account for multiple testing in the primary hypothesis tests.

Secondary outcomes will include measures such as burnout scores, task load, and professional fulfillment. These outcomes will be analyzed using similar statistical approaches, including comparisons of Vendor A and Vendor B to the control group, as well as an exploration of the relative impacts of Vendor A and Vendor B.

Our economic evaluations will entail a cost consequence analysis and a budget impact analysis.(24, 25) The cost-consequence analysis will separately present costs and outcomes associated with each intervention arm compared to the control arm. This analysis will show incremental costs as well as incremental benefits of each ambient AI scribe tool compared to the control group. The budget impact analysis will estimate the financial impact, from the perspective of the health system, of varying levels of implementation of each ambient AI scribe tool and project any potential gains in revenues through improvement in physician productivity and documentation efficiency.(24)

We will perform pre-specified secondary subgroup analyses to investigate heterogeneous treatment effects by introducing interactions between primary model terms and potential effect modifiers. Specifically, these analyses will evaluate whether weekly clinic days, provider sex, and the proportion of visits from new vs. established patients moderate the intervention effects. These analyses will be conducted to meet out hypothesized differences in effects due to higher baseline burnout rates in women, an

246 anticipated bigger dose response in physicians with more visits per week, and higher
247 utilization for new patients where previous note text cannot be copied forward.

249 *Sample Size*

250 The sample size for this study is constrained by exclusion criteria and a contractual
251 limitation allowing at most 100 concurrent users of any given tool. A sample size of at least
252 79 providers per condition will provide 80% power to detect effect sizes as small as 0.50
253 standard deviations, assuming a two-sample t-test and a two-sided 0.025 significance
254 level. This is a conservative simplification of the planned linear mixed-effects regression
255 analysis.

256 Using pre-study data from a comparable timeframe, the estimated log-scale
257 standard deviation of change for the time in notes per note metric is approximately 0.31.
258 This design will be powered to detect a 15.5% relative improvement in the primary
259 outcome. Given a baseline-year geometric mean time in notes per note of approximately
260 4m 43s, this corresponds to an absolute difference of roughly 44 seconds. This effect size
261 falls within the range of those seen in the literature(17).

263 **Discussion**

264 Our study aims to evaluate the impact of ambient AI scribes on EHR utilization,
265 physician burnout, and billing capture in the ambulatory setting through a three-arm RCT
266 over a 2-month period. The study is the largest RCT of ambient AI scribes to date and first to

include multiple vendors arms and a control arm. Results from this study will quantify how ambient AI scribes improve EHR documentation efficiency, affect burnout, and change physician engagement within a short intervention period. Additionally, this study evaluates physician-reported effects on note accuracy, patient safety, equity, and other unintended consequences, such as hallucinations or errors in text generation, addressing recognized limitations of LLM-based technologies.(26, 27) Similarly, understanding the economic impact of these tools, including costs related to physician adaptation and workflow changes, is essential to determine their value in real-world implementation.(18)

There are several limitations to this study. The relatively short duration of the intervention may not adequately capture the long-term effects of ambient AI scribes. Participants may require additional time to adapt to the ambient scribe technology, potentially delaying full proficiency. Since clinic schedules range from half-day to full-day sessions, the time needed for integration is likely to vary across participants. To address this variability, randomization was stratified by clinic schedules to ensure balanced adaptation periods among groups. Another potential limitation is that participation was voluntary, introducing the possibility of selection bias, as the findings may not fully generalize to all physicians. Furthermore, this study was conducted within a single academic health system using a specific EHR platform, which may limit the applicability of the results to other healthcare settings or systems.

This protocol does not evaluate the impact of ambient AI scribes on patient experience, satisfaction, or trust. While early studies suggest generally positive patient experiences with ambient AI scribes, existing studies have been limited by insufficient

power to draw definitive conclusions.(14, 15) Future research should explore not only patient satisfaction but also trust and perceptions of AI use in clinical settings, as these factors are crucial to widespread adoption. Looking ahead, we do hope to answer some if not all these questions in a larger roll out of the ambient AI scribe across the entire UCLA Health System.

This study may be subject to several potential biases. First, as participants will not be blinded to their assignment. The reliance on self-reported surveys for key secondary outcomes, such as physician satisfaction and burnout, introduces subjectivity that may not fully reflect the impact of the scribe tools. This is similar to previously published study by Owens et al., where patient perceptions varied based on their awareness of ambient scribe technology.(15) Lastly, encounters in languages other than English are excluded from the study, which limits the generalizability of our findings.

To our knowledge, this is the first RCT comparing two ambient AI scribe vendors. These findings will provide insights into their impact on efficiency, productivity and physician well-being in ambulatory care settings.

Trial Status

This study was reviewed by the UCLA Institutional Review Board (IRB-24-5425). Due to the complexity of vendor contract negotiations and the operational nature of the initiative, the timeline for finalizing agreements and starting the trial necessitated rapid implementation. The primary purpose of the initiative was to improve operations at the University of California, Los Angeles (UCLA) Health. Randomization was employed to

address secular trends, seasonal and holiday effects in December, and other factors confounding the relationship between exposure to the AI tools and the outcomes. Recognizing the value of rigorously studying the intervention's impact, the UCLA team opted to register the initiative despite its operational origin. As this initiative was operational rather than research-driven, it was essential to execute it within a compressed timeframe.

Recruitment for this study began on October 1, 2024, the trial intervention period commenced on November 4, 2024, and the trial period concluded on January 3, 2025. Due to the expedited timeline, the trial was completed before receiving registration approval, which was subsequently granted on ClinicalTrials.gov under the identifier NCT06792890, version 1.0, dated January 27, 2025 (<https://clinicaltrials.gov/study/NCT06792890>).

Declarations

Ethics Approval and Consent to Participate

This study was reviewed and approved by the UCLA Institutional Review Board (IRB-24-5425). Written informed consent to participate was obtained from all eligible participants prior to enrollment. All participants were informed about the study procedures, including data collection, and their right to withdraw at any time without penalty. No post-trial care or compensation is expected, as these physician level interventions are unlikely to cause harm to participants.

Consent for Publication

Not applicable. This manuscript does not contain any individual participant data requiring consent for publication.

Availability of Data and Materials

The datasets used and/or analyzed during this study will be available from the corresponding author upon reasonable request. Data sharing is subject to UCLA Health data access policies to ensure compliance with privacy and ethical standards.

Competing Interests

The authors declare that they have no competing interests related to this study. Ya-Chen Tina Shih serves on the OncoCollective Advisory Board for Sanofi Inc., unrelated to the content of this manuscript.

Funding

This study was funded by the UCLA Department of Medicine, with additional support from by the NIH/NIA award R01AG070017-01. JM was also supported by a NIH/NIA Beeson Emerging Leaders in Aging Research Career Development Award (Grant K76AG064392-01A1). CS is supported by the NIH/NIA Midcareer Award in Patient-Oriented Aging Research (Grant 1K24AG047899) and the NIH/NCATS UCLA Clinical and Translational Science Institute (CTSI) (Grant UL1TR001881, PI: Dubinett).

354 *Authors' Contributions*

355 All authors contributed to the conception and design of this trial. ATC and WT are equal
356 contributors.

357

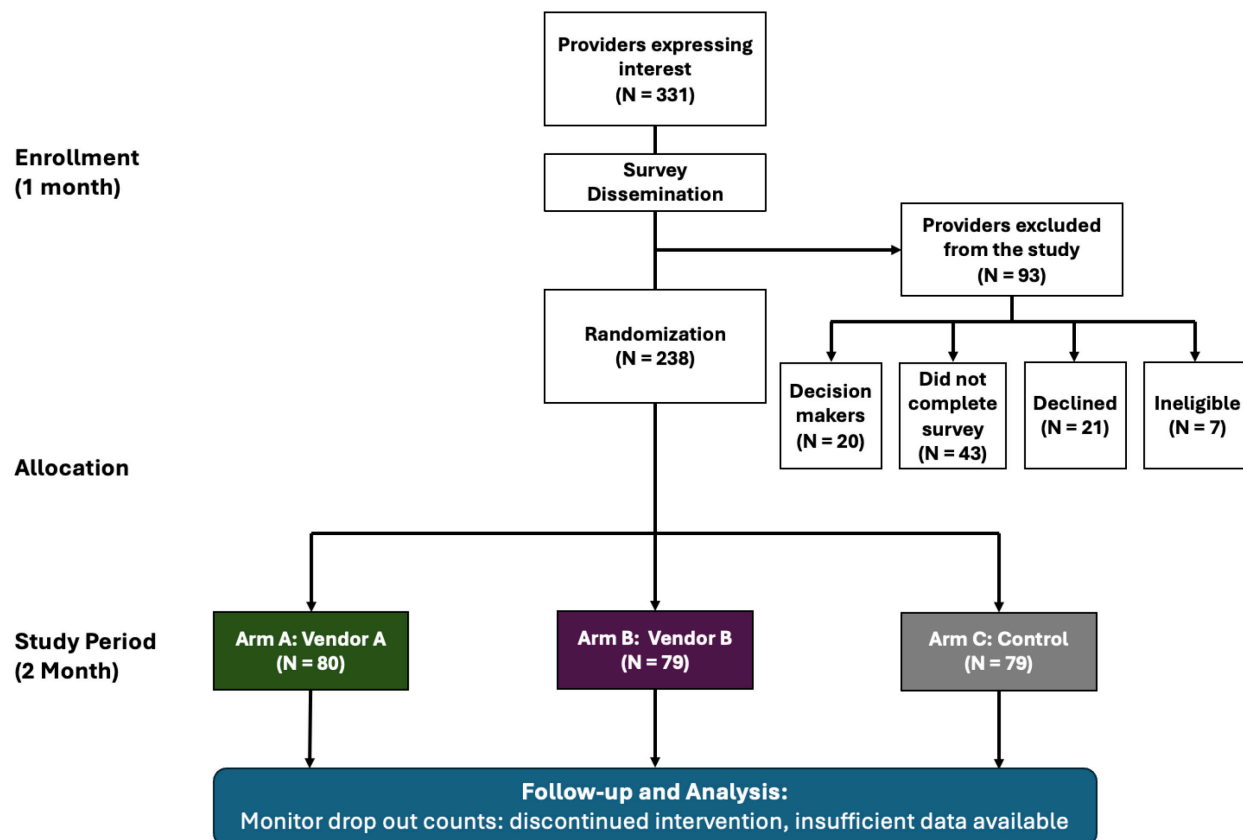
358 *Acknowledgments*

359 We thank the UCLA Health Information Technology team for their technical support during
360 the study and the participating physicians for their time and commitment to this research.

361 We acknowledge and appreciate the work of Chad Wes Villaflores, UVAS Senior Data
362 Scientist, for leading the ClinicalTrials.gov submission, and Artem Romanov for his
363 contributions to the IRB submission.

364

365 **Figures and Tables**



367

368 **Figure 1 – CONSORT-AI** flow diagram illustrating the recruitment, randomization, and
 369 allocation of study participants. A total of 331 providers expressed interest, of which 93
 370 were excluded due to ineligibility, declining participation, incomplete surveys, or decision-
 371 making roles. The remaining 238 participants were randomized into three arms: Vendor A (n
 372 = 80), Vendor B (n = 79), and Control (n = 79). The study spans a two-month intervention
 373 period, with follow-up and analysis monitoring dropout counts due to discontinued
 374 intervention or insufficient data availability

375

Study Period	Enrollment	Allocation	Study Period
Timepoint	-1 month	- month	2 months
	Oct 2024	Nov 4, 2024	Nov 4, 2024 – Jan 3, 2025
Enrollment			
Eligibility Screen	X		
Informed Consent	X		
Randomization		X	
Interventions			
Vendor A or B			X
Control (Arm C)			
Provider-based Assessments^Φ			
Time in Notes per note (primary) [†]		X	X
Pajama time			X
After-hours (7 PM– 7 AM)			X
Mini-Z 2.0 Burnout Score	X		X
Professional Fulfillment Index	X		X
NASA Task Load	X		X
Physician Satisfaction	X		X
Net Promoter Score			X
Appointment Turnover Time			X
Relative-Value Units (RVU) per clinic			X

^Φ Change from baseline

[†] The primary outcome for 'Time in Notes per Note' is the second month (Dec 3, 2024 – Jan 3, 2025), though data from the entire study period are analyzed to evaluate period effects.

Table 1 – SPIRIT-AI Schedule of enrollment, interventions, and assessments outlining key study activities, including eligibility screening, consent, randomization, interventions, and outcome assessments. Primary and secondary measures, such as time in notes per note, burnout scores, task load, and productivity metrics, are collected at specified intervals during the enrollment, allocation, and study periods (October–December 2024)

References

1. Peccoraro LA, Kaplan CA, Pietrzak RH, Charney DS, Ripp JA. The impact of time spent on the electronic health record after work and of clerical work on burnout among clinical faculty. *Journal of the American Medical Informatics Association*. 2021/04/23;28(5).
2. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, Rossetti SC. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal of the American Medical Informatics Association*. 2021/04/23;28(5).
3. Lou SS, Lew D, Harford DR, Lu C, Evanoff BA, Duncan JG, Kannampallil T. Temporal Associations Between EHR-Derived Workload, Burnout, and Errors: a Prospective Cohort Study. *Journal of General Internal Medicine*. 2022/07;37(9).
4. Guille C, Sen S. Burnout, Depression, and Diminished Well-Being among Physicians. *New England Journal of Medicine*. 2024-10-24;391(16).
5. Sittig DF, Singh H. Recommendations to Ensure Safety of AI in Real-World Clinical Care. *JAMA*. 2024.
6. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*. 2016-09-06;165(11).
7. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. *JAMA Network Open*. 2022/09/01;5(9).

- 405 8. Mishra P, Kiang JC, Grant RW. Association of Medical Scribes in Primary Care With
406 Physician Workflow and Patient Experience. JAMA Internal Medicine. 2018/11/01;178(11).
- 407 9. Bates DW, Landman AB. Use of Medical Scribes to Reduce Documentation Burden.
408 JAMA Internal Medicine. 2018/11/01;178(11).
- 409 10. Heckman J, Mukamal KJ, Christensen A, Reynolds EE, Heckman J, Mukamal KJ, et al.
410 Medical Scribes, Provider and Patient Experience, and Patient Throughput: a Trial in an
411 Academic General Internal Medicine Practice. Journal of General Internal Medicine 2019
412 35:3. 2019-12-05;35(3).
- 413 11. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical
414 documentation from 1990 to 2018: a systematic review. Journal of the American Medical
415 Informatics Association. 2019/04/01;26(4).
- 416 12. Liu T-L, Hetherington TC, Stephens C, McWilliams A, Dharod A, Carroll T, Cleveland
417 JA. AI-Powered Clinical Documentation and Clinicians' EHR Experience. JAMA Network
418 Open. 2024/09/03;7(9).
- 419 13. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, et al. The
420 impact of nuance DAX ambient listening AI documentation: a cohort study. Journal of the
421 American Medical Informatics Association. 2024/04/03;31(4).
- 422 14. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient
423 Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM
424 Catalyst Innovations in Care Delivery. 2024-02-21;5(3):CAT. 23.0404.

- 425 15. Owens LM, Wilda J, Westendorp J, Grifka R, Fletcher J. Effect of Ambient Voice
426 Technology, Natural Language Processing, and Artificial Intelligence on the Patient-
427 Physician Relationship. *Applied clinical informatics*. 2024 Aug;15(4).
- 428 16. Cao DY, Silkey JR, Decker MC, Wanat KA. Artificial intelligence-driven digital scribes
429 in clinical documentation: Pilot study assessing the impact on dermatologist workflow and
430 patient encounters. *JAAD International*. 2024/06/01;15.
- 431 17. Rotenstein L, Melnick ER, Iannaccone C, Zhang J, Mugal A, Lipsitz SR, et al. Virtual
432 Scribes and Physician Time Spent on Electronic Health Records. *JAMA Network Open*.
433 2024/05/01;7(5).
- 434 18. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the Economic
435 Value of Clinical Artificial Intelligence: Challenges and Opportunities. *Value in Health*.
436 2022/03/01;25(3).
- 437 19. McCoy LG, Manrai AK, Rodman A. Large Language Models and the Degradation of
438 the Medical Record. *The New England journal of medicine*. 10/31/2024;391(17).
- 439 20. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, Cruz Rivera S, et al.
440 Guidelines for clinical trial protocols for interventions involving artificial intelligence: the
441 SPIRIT-AI extension. *Nature Medicine* 2020 26:9. 2020-09-09;26(9).
- 442 21. Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, et al. Artificial
443 Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Network Open*.
444 2024/03/04;7(3).
- 445 22. Trockel M, Bohman B, Lesure E, Hamidi MS, Welle D, Roberts L, et al. A Brief
446 Instrument to Assess Both Burnout and Professional Fulfillment in Physicians: Reliability

447 and Validity, Including Correlation with Self-Reported Medical Errors, in a Sample of
448 Resident and Practicing Physicians. *Academic Psychiatry* 2017 42:1. 2017-12-01;42(1).

449 23. Linzer M, McLoughlin C, Poplau S, Goelz E, Brown R, Sinsky C, et al. The Mini Z
450 Worklife and Burnout Reduction Instrument: Psychometrics and Clinical Implications.
451 *Journal of General Internal Medicine* 2021 37:11. 2022-01-19;37(11).

452 24. Sullivan SD, Mauskopf JA, Augustovski F, Caro JJ, Lee KM, Minchin M, et al. Budget
453 Impact Analysis—Principles of Good Practice: Report of the ISPOR 2012 Budget Impact
454 Analysis Good Practice II Task Force. *Value in Health*. 2014/01/01;17(1).

455 25. Mauskopf JA, Paul JE, Grant DM, Stergachis A, Mauskopf JA, Paul JE, et al. The Role
456 of Cost—Consequence Analysis in Healthcare Decision—Making. *Pharmacoeconomics*
457 1998 13:3. 2012-10-18;13(3).

458 26. Koenecke A, Gyeong Choi, Anna Seo, Mei, Katelyn X., Schellmann, Hilke, Sloane,
459 Mona, Koenecke, Allison, Gyeong Choi, Anna Seo, Mei, Katelyn X., Schellmann, Hilke, Sloane,
460 Mona. Careless Whisper: Speech-to-Text Hallucination Harms | Proceedings of the 2024
461 ACM Conference on Fairness, Accountability, and Transparency. The 2024 ACM
462 Conference on Fairness, Accountability, and Transparency. 2024.

463 27. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine
464 Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine*.
465 11/01/2018;178(11).

466