A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being

Majid Afshar, MD, MS
Associate Professor
Director, Learning Health Systems
Departments of Medicine and Biostatistics & Medical Informatics
University of Wisconsin-Madison

Mary Ryan Baumann, PhD
Assistant Professor
Departments of Population Health Sciences and Biostatistics & Medical Informatics
University of Wisconsin-Madison

UW Health | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Disclosures

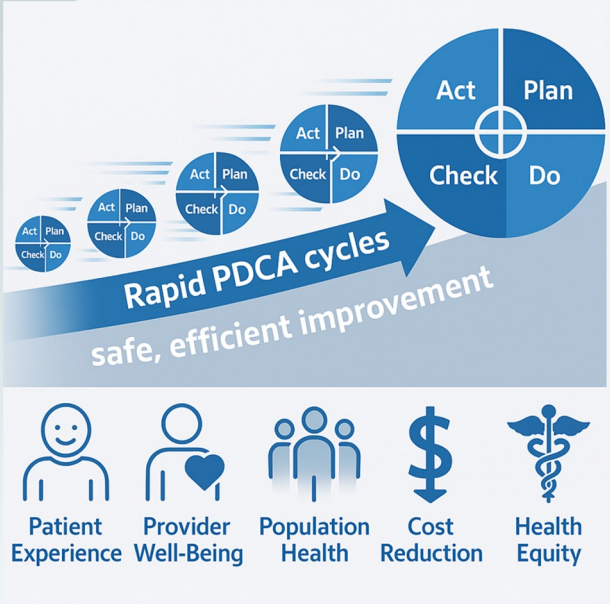
This work was supported by funding from the University of Wisconsin Hospital and Clinics and the National Institute of Health Clinical and Translational Science Award (NIH/NCATS UL1TR002737). No funding was provided by the AI software company and all licenses of the software were procured as a vendor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

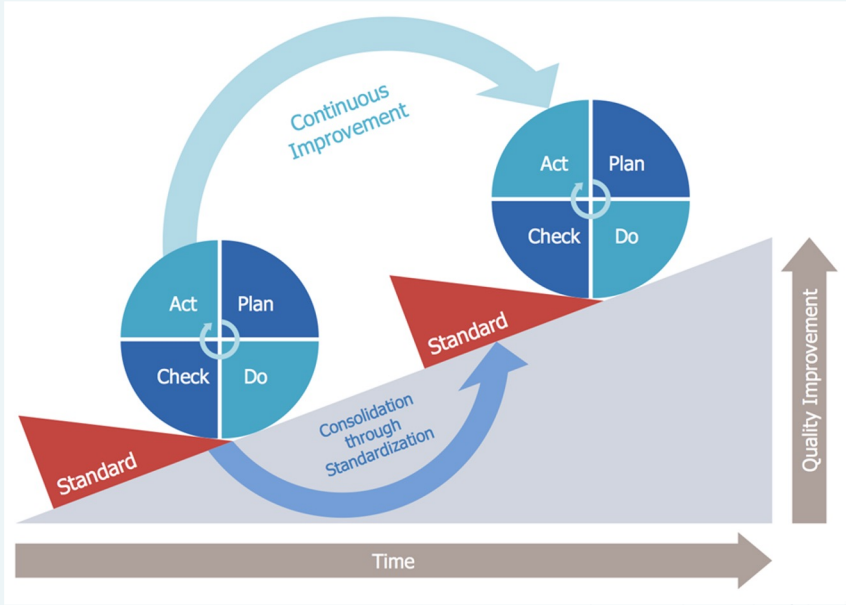
Learning Health System LHS-CP

Data-Driven, Implementation-focused Learning Health System

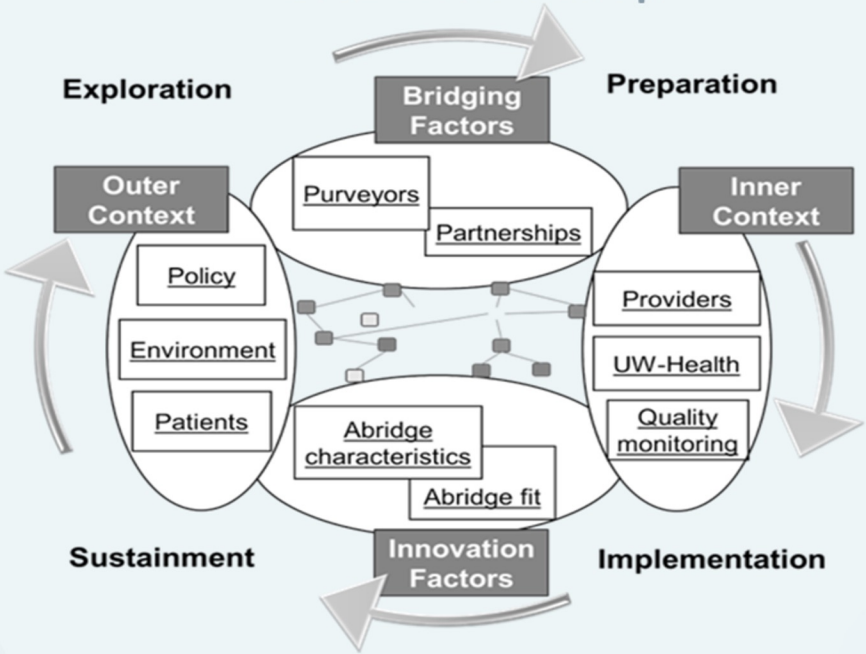
UWHealth | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Learn from what we do, and do what we learn





Systems Engineering/Implementation Science Framework with Pilot Group



EPIS Model at UW Health applied to Ambient Listening project to improve fidelity of AI Software

UWHealth | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

Pragmatic Trial Operations (PTOps) Playbook

- **Clinical Trial Protocol and IRB Application**
- **Governance**
  - Project charter, RIDAC, Escalation Matrix, Consent forms, ct.gov, committee approvals
- **User Experience**
  - REDCap surveys, interview guides and summaries, framework
  - Implementation Scientist/D&I Launchpad
- **Technical Integration and Onboarding**
  - User onboarding and training user guide, FAQ, patient flyer, technical specs
- **Analytics and Data Dashboard Monitoring**
  - Monitoring and Drift Detection
  - LLM Coder
  - PDSQI-9 LLM-as-a-Judge documentation quality
  - EHR Secondary outcomes/process measures
  - Informatics Services and Biostatistics
- **Clinical Trial**
  - Data Dictionary, REDCap Codebook, Statistical Analysis Plan and Code, Randomization code,
  - De Identified Primary Outcome Data
  - Clinical Trialist

UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Ambient Listening: The Experience Problem

Manual entry

Transcription

Front End Speech Recognition

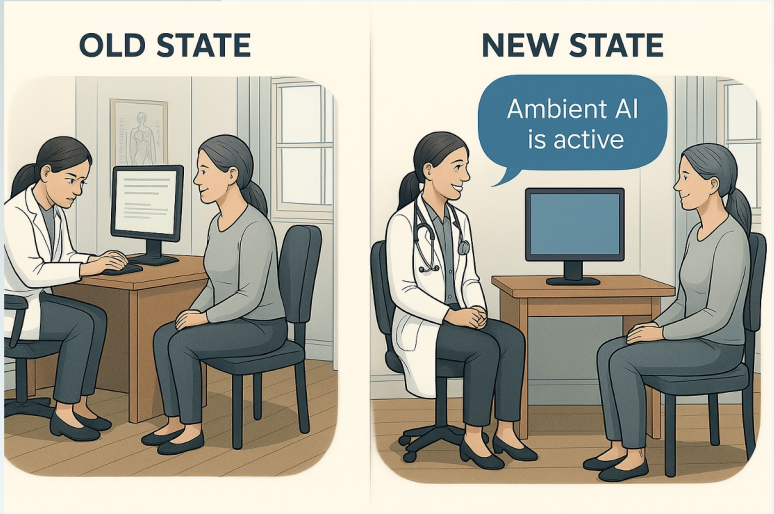
Scribe

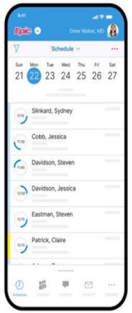
Doctor **Patient**

Courtesy of Joel Gordon, MD (CMIO at UW Health)

UWHealth | School of Medicine and Public Health UNIVERSITY OF WISCONSIN-MADISON

Ambient AI Software Intervention



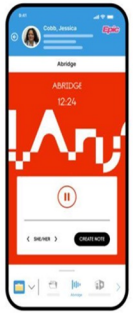


STEP 1

Select a patient from your schedule within Haiku and click "Abridge" at the bottom of your screen.

STEP 2

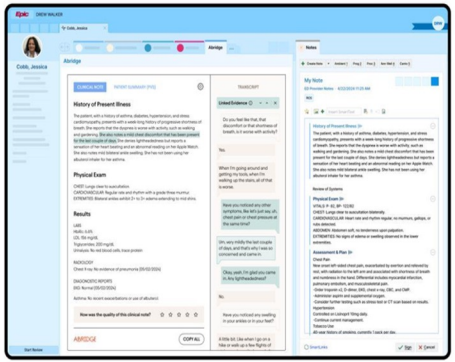
Tap the red record button and conduct the patient encounter.



STEP 3

When finished recording, tap "Create Note."

→ A draft note is created via audio recording and AI summarization.



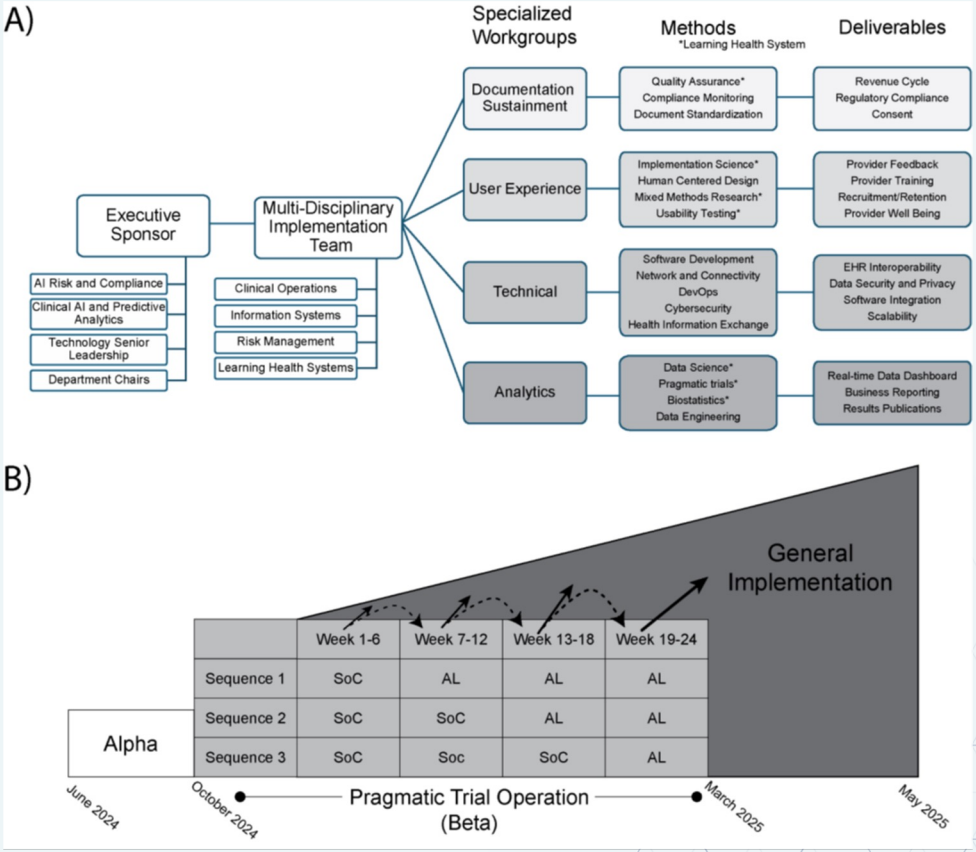
STEP 4

Access the Abridge tab within Hyperspace to edit and verify your note. Finalize and sign the note.

STEP 5

Parts of the edited note are pulled into Epic using dot phrases. These dot phrases are embedded into our standard note template.

Governance and Workflow



Trial Design

Broad Eligibility Criteria

Inclusion Criteria	Exclusion Criteria
Healthcare Provider in the UW Health Outpatient Clinic Setting	Planned leave in 6 weeks following randomization
Willingness to engage and use ambient technology	Not registered onto Epic's mobile Haiku system for access
Individuals at least 18 years of age	User of existing medical scribe and unwilling to discontinue
English and/or Spanish speaking	
Complete training and in-servicing of the tool	
Providing outpatient care to no less than 20 encounters on a weekly average	
Owns an Apple mobile device, as the software is accessible only on this platform	

Primary Outcomes: Stanford Professional Fulfillment Index

- Co-primary well-being outcomes from validated Stanford Professional Fulfillment Index (PFI)
  - Professional Fulfillment subscale (1-5)
  - Burnout composite (1-5): Work exhaustion subscale + Professional disengagement subscale
- Pragmatism
  - PFI previously integrated into routine data collection by health system Chief Quality Officer
- Synergy
  - Consultation between Research and Operations teams to ID operationally meaningful outcomes (quintuple aims)
- Historic data informed minimum number of providers needed for trial

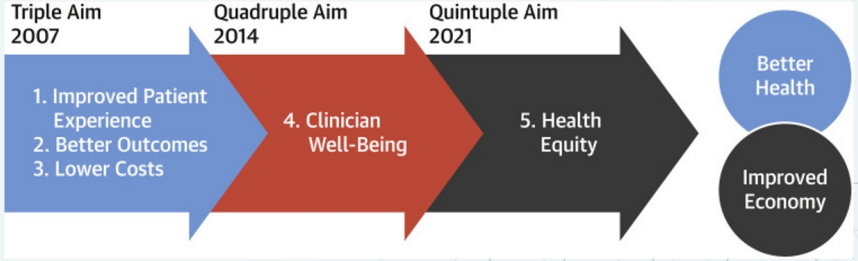


Figure via Itchhaporia (2021). "The Evolution of the Quintuple Aim". J Am Coll Cardiol.

Outcome Measures

2 co-primary endpoints: overall physician well-being

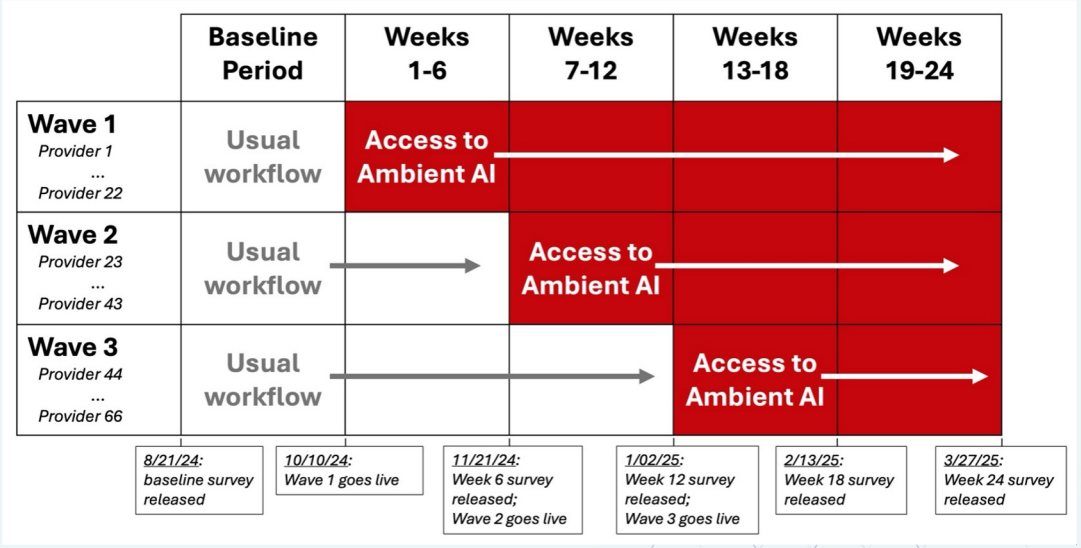
- Professional Fulfillment subscale
- Burnout composite

10 secondary endpoints: documentation, specific well-being

- Time in Notes
- Work Outside of Work
- Encounters with note closeout before next patient encounter
- Encounters with note closeout same day
- Patient follow up within 2 weeks
- Task Load
- Negative Impact of Work on Meaningful Relationships
- Meaningfulness of Work on Interpersonal Relationships
- Control Over Schedule
- Trust in AI

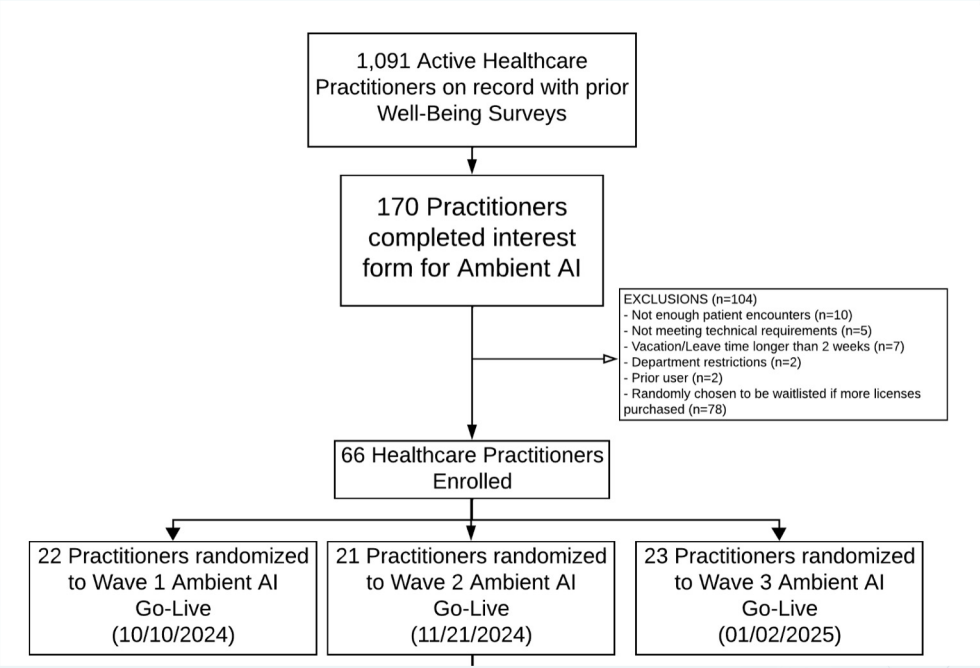
Trial Design Scheme

- 1:1:1 individual randomization
- Stratified by specialty (*Family Medicine, Internal Medicine, Pediatrics, Other Specialties*)
- Within-provider ICC=0.65 based on reported test-retest reliability
- 85% power for Cohen's d of 0.44 (0.025 two-sided type I error) – clinically meaningful based on prior literature



UWHealth | School of Medicine and Public Health, University of Wisconsin-Madison

Reality: Pre-Go-Live CONSORT Flow



Analysis

- Linear mixed effect models
  - Provider-level random intercept
  - Adjustment for categorical time-period effects
  - Adjustment for randomization stratification
  - Exploratory analyses: time-varying treatment effect (treatment-by-time interaction)
- Type I error control
  - Co-primary outcomes: even split of 0.05 error
  - Secondary outcomes: Benjamini Hochberg adjustment

Data Capture & Construction

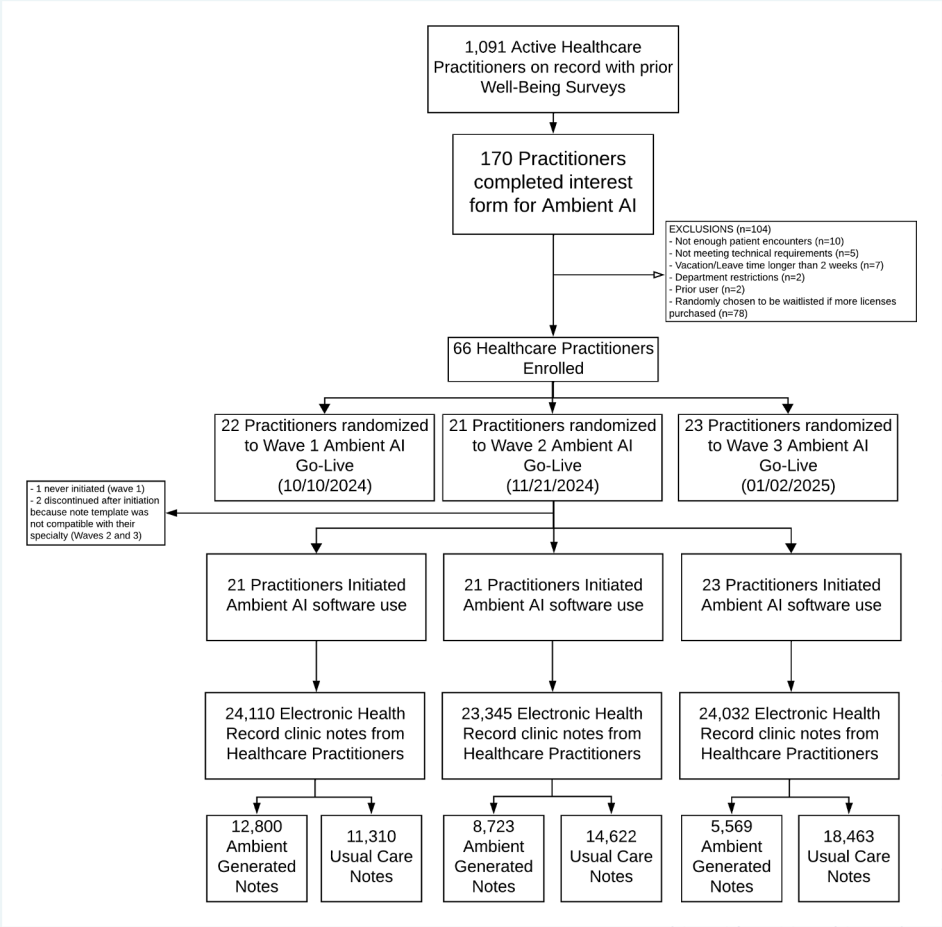
- Multiple data sources & observation levels
  - REDCap survey data (period-level)
  - EHR data warehouse (encounter note level)
  - AI Software data (encounter experience-level)
- Partnership with Enterprise Analytics team (health system operations) to construct & deliver meaningful documentation metrics (encounter note- and day-level)
- Data harmonization – not as easy as it sounds!
  - Capture of note changes
  - Normalization to 8-hour workday

Main Trial Results

UWHealth | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Full Trial CONSORT Flow: intention to treat analysis

- Note-weighted median utilization: 71%
- Survey completion: 99.7%
- Patient consent: 99.92% (n=22 declined)



UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Provider Baseline Characteristics

	Wave 1 (N=22)	Wave 2 (N=21)	Wave 3 (N=23)	Overall (N=66)
Female Sex, *n (%)*	19 (86.4%)	15 (71.4%)	18 (78.3%)	52 (78.8%)
Age, *median [IQR]*	40.5 [38.3, 46.8]	48.0 [39.0, 55.0]	42.0 [35.5, 47.0]	42.5 [36.3, 49.0]
Race/Ethnicity, *n (%)*				
AI/AN	0	0	0	0
Non-SE Asian	1 (4.5%)	0	0	1 (1.5%)
AA/Black	0	0	0	0
Hispanic/Latino	0	1 (4.8%)	0	1 (1.5%)
ME/NA	0	0	0	0
NH/PI	0	0	0	0
SE Asian	1 (4.5%)	0	0	1 (1.5%)
White	18 (81.8%)	19 (90.5%)	22 (95.7%)	59 (89.4%)
Multiple	2 (9.1%)	1 (4.8%)	0	3 (4.5%)
Missing	*0*	*0*	*1 (4.3%)*	*1 (1.5%)*
Years in practice, *median [IQR]*	15.0 [10.5, 16.0]	13.0 [9.5, 26.0]	14.0 [9.0, 21.0]	14.0 [9.0, 20.0]
Missing (%)	*1 (4.5%)*	*0*	*0*	*1 (1.5%)*

	Wave 1 (N=22)	Wave 2 (N=21)	Wave 3 (N=23)	Overall (N=66)
Provider Type, *n (%)*				
NP	2 (9.1%)	1 (4.8%)	2 (8.7%)	5 (7.6%)
PA	5 (22.7%)	6 (28.6%)	2 (8.7%)	13 (19.7%)
Physician	15 (68.2%)	14 (66.7%)	19 (82.6%)	48 (72.7%)
Randomization speciality stratum, *n (%)*				
Family Medicine	10 (45.5%)	9 (42.9%)	11 (47.8%)	30 (45.5%)
Internal Medicine	6 (27.3%)	6 (28.6%)	6 (26.1%)	18 (27.3%)
Pediatrics/Adolescent Medicine	2 (9.1%)	2 (9.5%)	2 (8.7%)	6 (9.1%)
Other Speciality[‡]	4 (18.2%)	4 (19.0%)	4 (17.4%)	12 (18.2%)
Patients seen in the clinic per week, *median [IQR]*	55 [36.0, 62.0]	41 [36.0, 57.0]	50 [35.5, 59.0]	50 [36, 60]
Missing, n (%)	*1 (4.5%)*	0	0	1 (1.5%)

UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Provider Baseline Characteristics

	Wave 1 (N=22)	Wave 2 (N=21)	Wave 3 (N=23)	Overall (N=66)
Pre-trial note taking method†, *n (%)*				
Manual typing	17 (77.3%)	19 (90.5%)	19 (82.6%)	55 (83.3%)
Templated	18 (81.8%)	19 (90.5%)	19 (82.6%)	56 (84.8%)
Dictation	2 (9.1%)	0	1 (4.3%)	3 (4.5%)
Scribes	3 (13.6%)	5 (23.8%)	4 (17.4%)	12 (18.2%)
Fluency Direct	13 (58.1%)	13 (61.9%)	13 (56.5%)	39 (59.1%)
Other (student)	0	1 (4.8%)	0	1 (1.5%)
PFI Professional Fulfillment (baseline), *mean (SD)*	3.66 (0.72)	3.56 (0.74)	3.08 (0.69)	3.42 (0.75)
PFI Burnout (baseline), *mean (SD)*	2.36 (0.80)	2.33 (0.65)	2.74 (0.56)	2.49 (0.96)

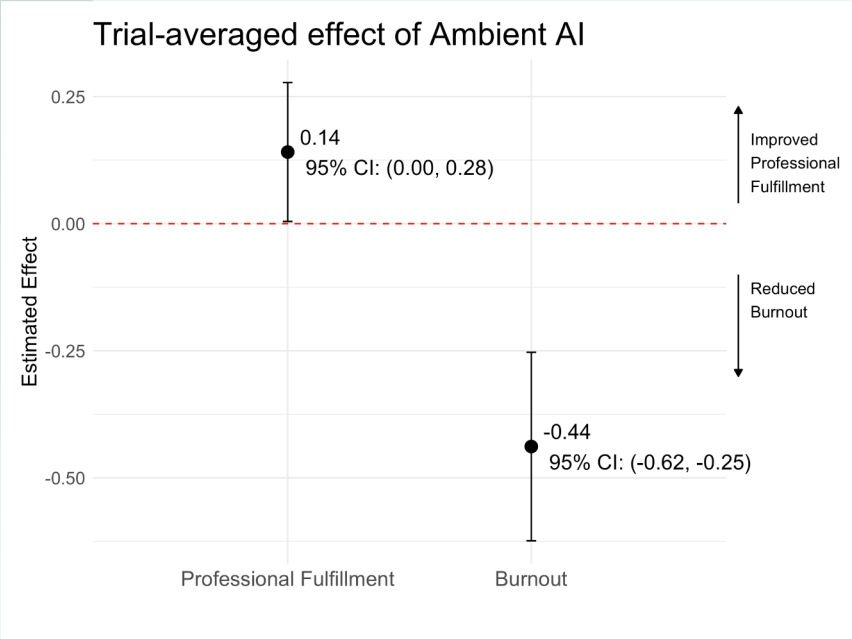
Patient Encounter Baseline Characteristics

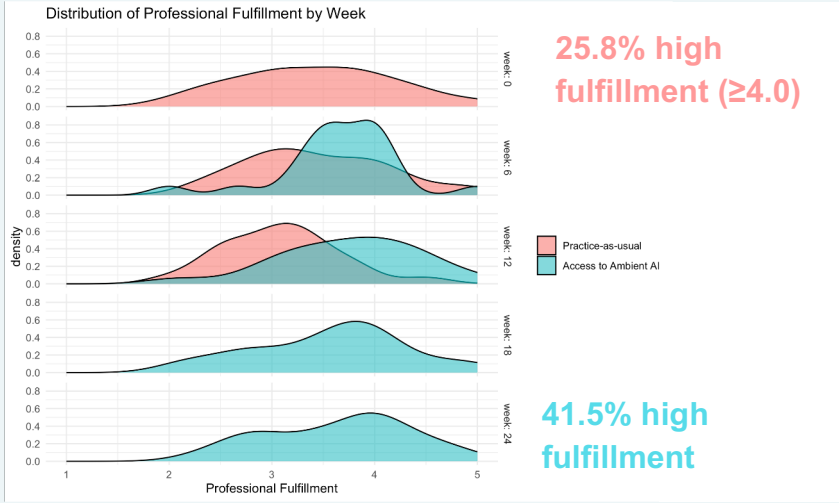
- 40.2% female
- Median age: 22 years (IQR: 27-67)
- 88.4% White, 93.4% non-Hispanic
- 98.1% English preferred language
- 92.1% conducted in-person

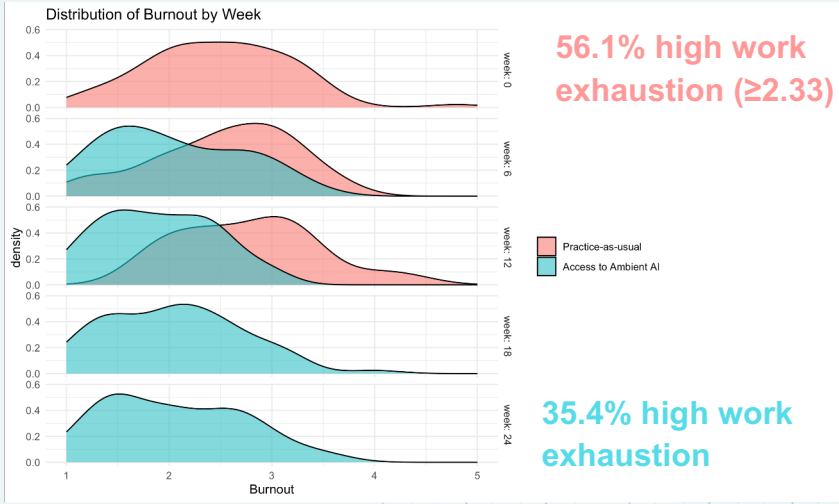
	Wave 1 Group (N=5,740)	Wave 2 Group (N=5,370)	Wave 3 Group (N=5,483)	Overall (N=16,593)
Female Sex, *n (%)*	2,247 (39.1%)	2,224 (41.4%)	2,197 (40.1%)	6,668 (40.2%)
Age at encounter, *median [IQR]*	50.0 [30.0, 67.0]	49.0 [28.0, 67.0]	48.0 [25.0, 67.0]	49 .0 [27.0, 67.0]
Race, *n (%)*				
Asian	204 (3.6%)	210 (3.9%)	205 (3.7%)	619 (3.7%)
AA/Black	327 (5.7%)	367 (6.8%)	263 (4.8%)	957 (5.8%)
AI/AN	42 (0.7%)	36 (0.7%)	32 (0.6%)	110 (0.7%)
NH/PI	2 (0.03%)	8 (0.15%)	4 (0.07%)	14 (0.08%)
White	5,087 (88.6%)	4,675 (87.1%)	4,914 (89.6%)	14,676 (88.4%)
Unknown	8 (0.1%)	10 (0.19%)	6 (0.1%)	24 (0.1%)
Decline/blank	70 (1.2%)	64 (1.2%)	59 (1.1%)	193 (1.1%)
Ethnicity, *n (%)*				
Hispanic/Latino	339 (5.9%)	329 (6.1%)	335 (6.1%)	1,003 (6.0%)
Not Hispanic/Latino	5,360 (93.4%)	5,014 (93.4%)	5,117 (93.3%)	15,491 (93.4%)
Unknown	5 (0.9%)	3 (0.06%)	0	8 (0.05%)
Decline/blank	36 (0.6%)	24 (0.4%)	31 (0.6%)	91 (0.5%)
Preferred language, *n (%)*				
American Sign Language	5 (0.09%)	2 (0.04%)	4 (0.07%)	11 (0.07%)
English	5,636 (98.2%)	5,251 (97.8%)	5,393 (98.4%)	16,280 (98.1%)
Hmong	3 (0.08%)	7 (0.1%)	5 (0.09%)	15 (0.09%)
Spanish	70 (1.2%)	79 (1.5%)	52 (0.9%)	201 (1.2%)
Other	25 (0.4%)	30 (0.6%)	29 (0.5%)	84 (0.5%)
Blank	1 (0.03%)	1 (0.02%)	0	2 (0.01%)
Insurance type, *n (%)*				
Blue Shield	755 (13.2%)	628 (11.7%)	722 (13.2%)	2,105 (12.7%)
Commercial/Commercial FFS	941 (16.4%)	968 (18.0%)	953 (17.4%)	2,862 (17.2%)
Medicaid/Medicaid MCO	483 (8.4%)	549 (10.2%)	498 (9.1%)	1,530 (9.2%)
Medicare	1,012 (17.6%)	1,012 (18.8%)	1,165 (21.2%)	3,189 (19.2%)
Medicare Advantage	647 (11.3%)	613 (11.4%)	491 (9 .0%)	1,751 (10.6%)
Quartz/Unity	1,711 (29.8%)	1,435 (26.8%)	1,499 (27.3%)	4,645 (28.0%)
Workers Comp/Other	18 (0.3%)	8 (0.1%)	10 (0.2%)	36 (0.2%)
None/Self-pay	173 (3.0%)	157 (2.9%)	145 (2.6%)	475 (2.9%)
Body Mass Index, *median [IQR]*	27.3 [22.7, 33.0]	27.6 [22.8, 33.3]	26.0 [21.2, 31.3]	27.0 [22.2, 32.5]
Missing, n (%)	*161 (2.8%)*	*348 (6.5%)*	*270 (4.9%)*	*779 (4.7%)*
First Systolic Blood Pressure, *median [IQR]*	121 [110, 133]	123 [112, 135]	121 [111, 134]	122 [111, 134]
Missing, n (%)	*894 (15.6%)*	*944 (17.6%)*	*2,091 (38.1%)*	*3,929 (23.7%)*
First Diastolic Blood Pressure, *median [IQR]*	76.0 [70.0, 82.0]	77.0 [71.0, 83.0]	76.0 [70.0, 82.0]	76.0 [70.0, 82.0]
Missing, n (%)	*894 (15.6%)*	*944 (17.6%)*	*2,091 (38.1%)*	*3,929 (23.7%)*
Number of medications prescribed, *median [IQR]*	1.0 [0, 2.0]	1.0 [0, 2.0]	1.0 [0, 2.0]	1.0 [0, 2.0]
Elixhauser score, *Mean (SD)*	0.12 (2.51)	0.47 (2.29)	0.27 (2.04)	0.28 (2.30)
Missing, n (%)	*5 (0.1%)*	*5 (0.1%)*	*5 (0.1%)*	*15 (0.1%)*

UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Primary Outcomes: Professional Fulfillment and Burnout

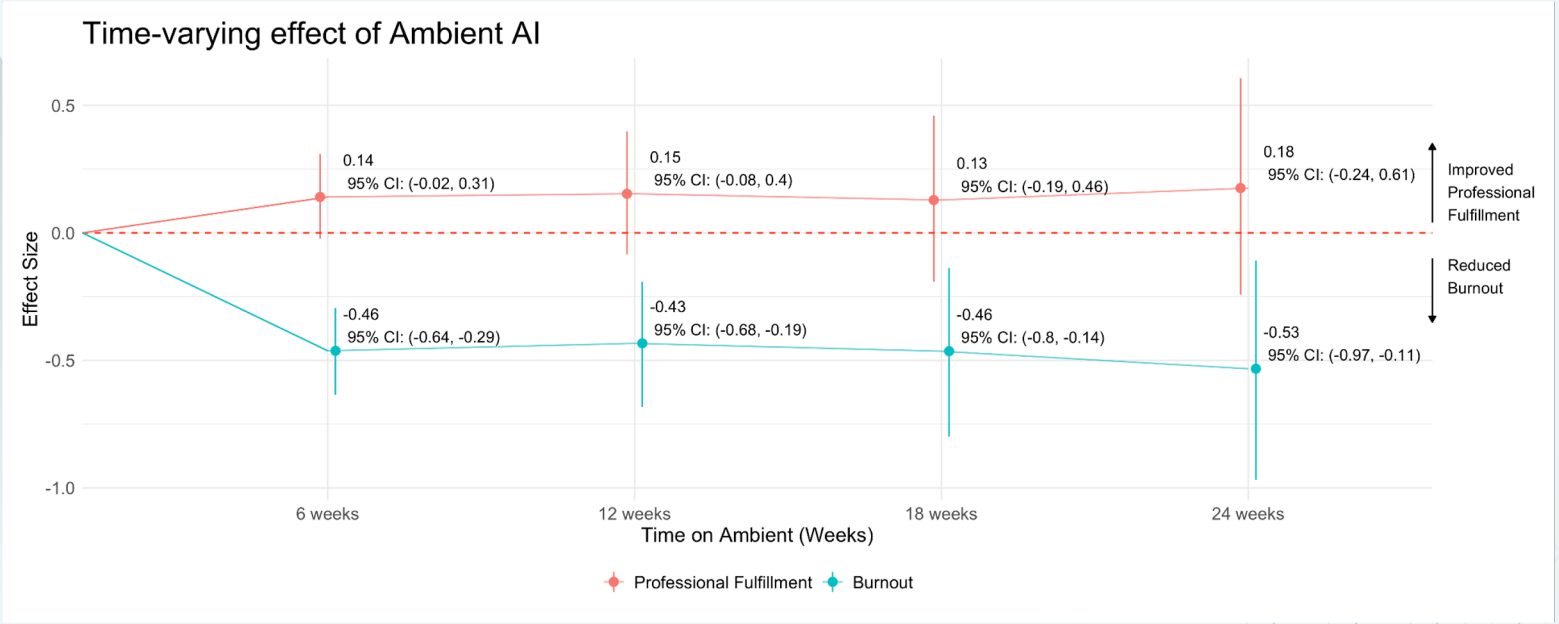






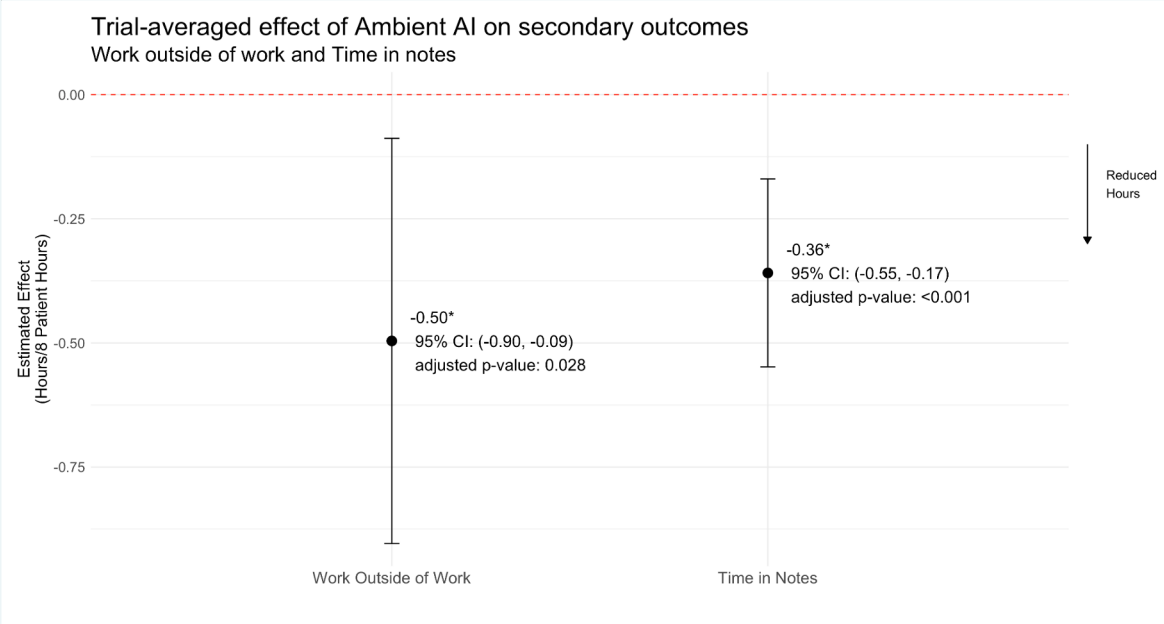
UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Primary Outcomes: (Lack of) Effect Variation Over Time



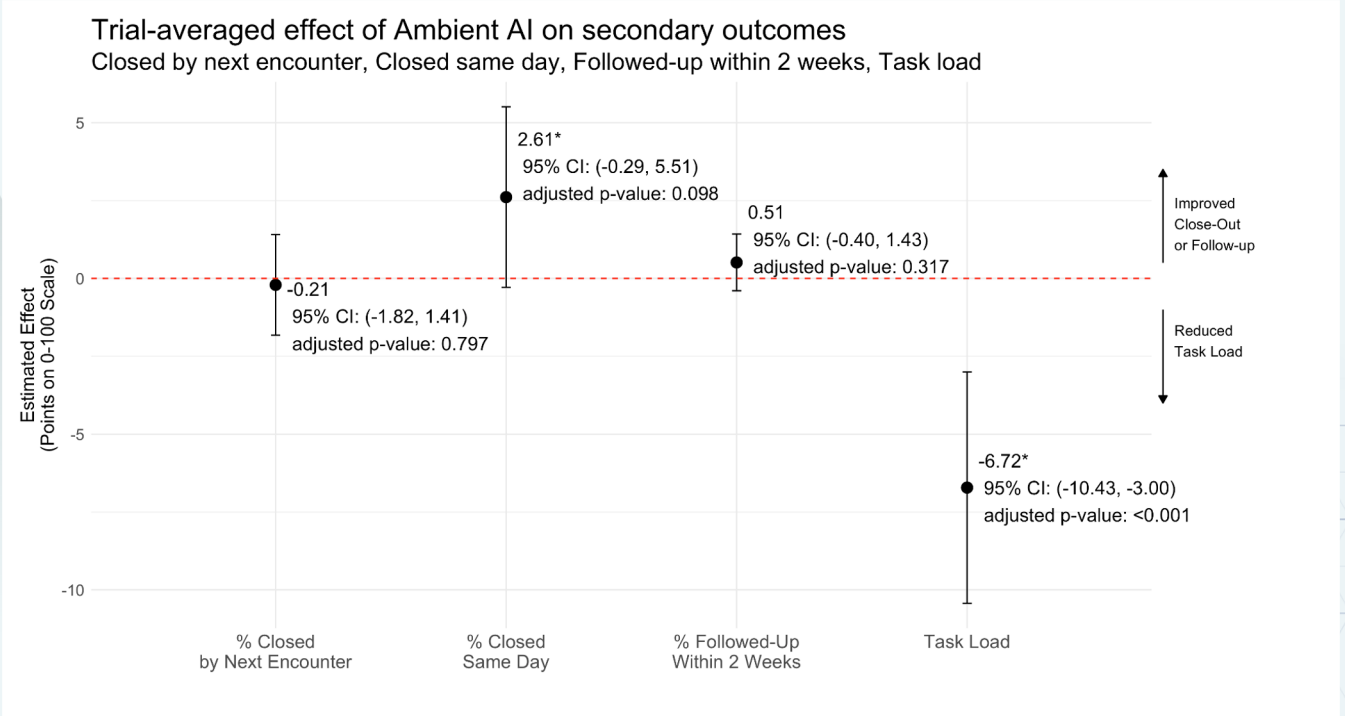
Secondary Outcomes: Documentation Time

- Outlier artifacts introduced during normalization of EHR usage metrics
- WoW
  - Exclude top 3% of observations
  - New effect size non-significant: –0.19 hours; 95% CI (–0.38, 0.00)
- Time in Notes
  - Exclude top 0.5% of observations
  - New effect size maintains statistical significance: –0.34 hours; 95% CI (–0.43, –0.25)

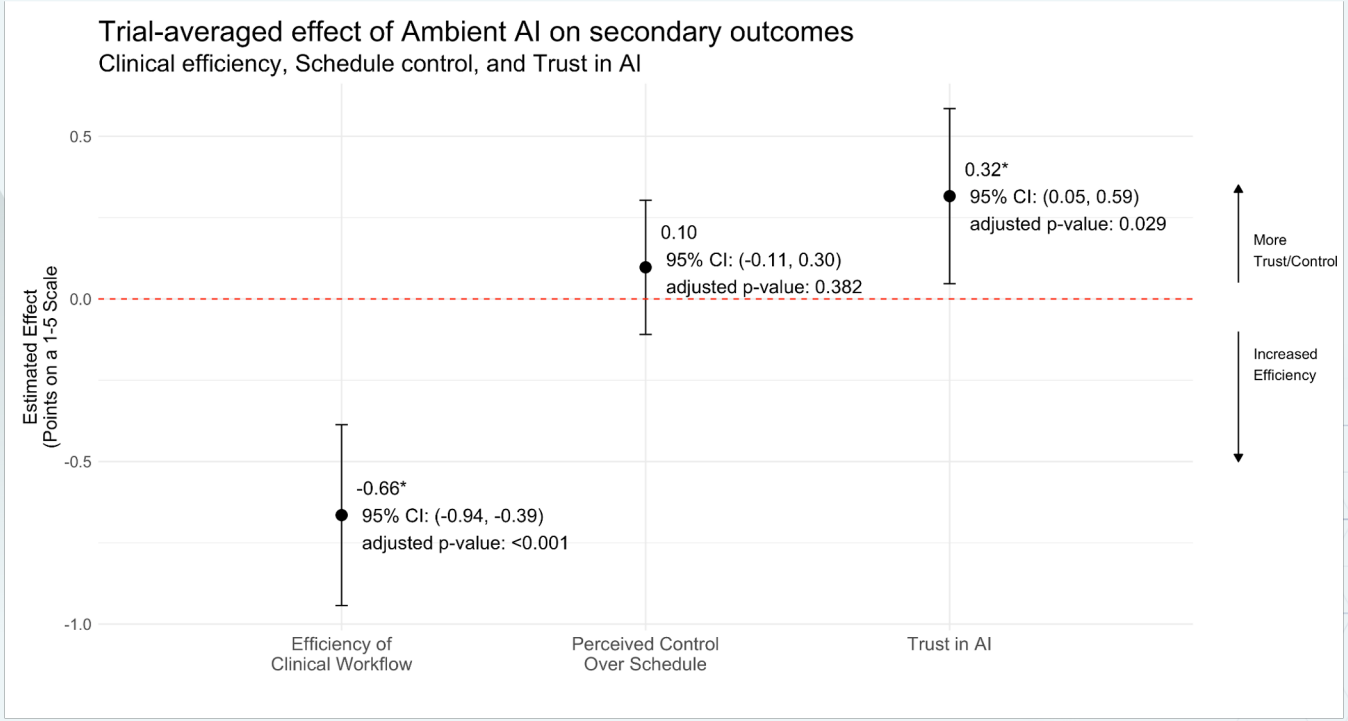


UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Secondary Outcomes: Close-Out and Task Load

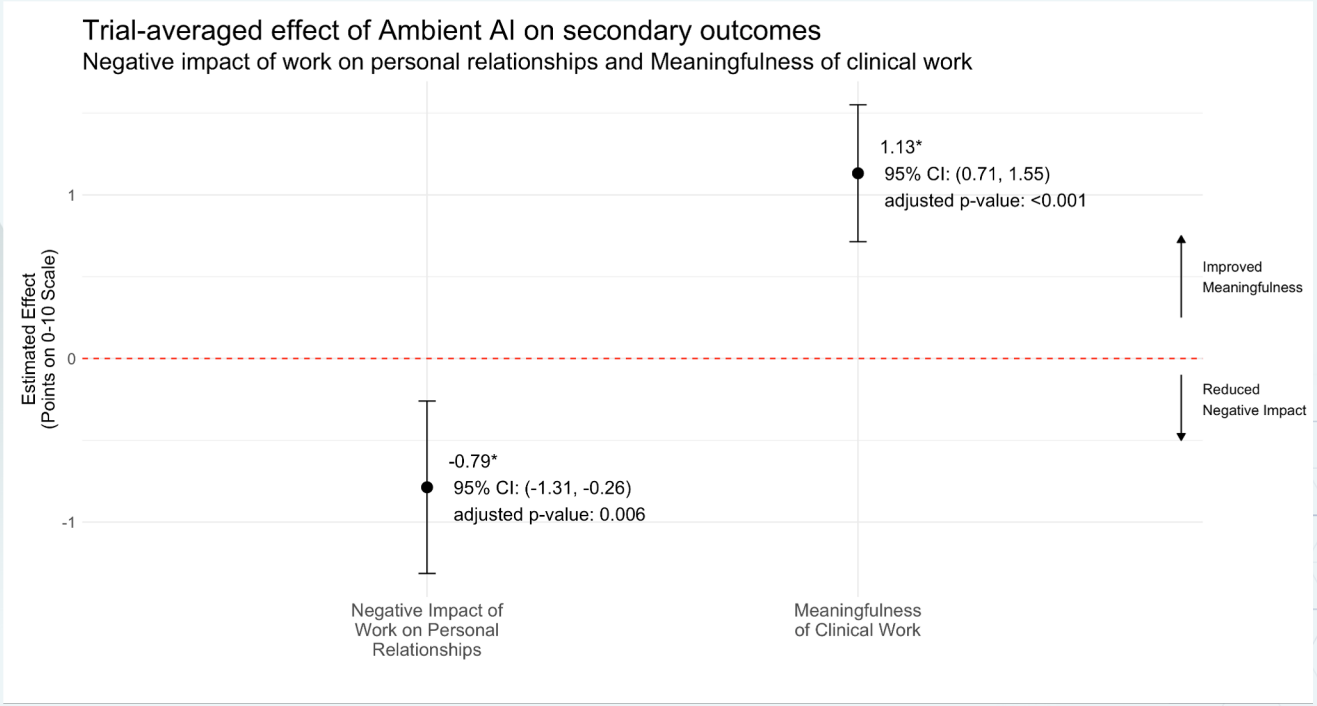


Secondary Outcomes: Perceptions of Efficiency and AI

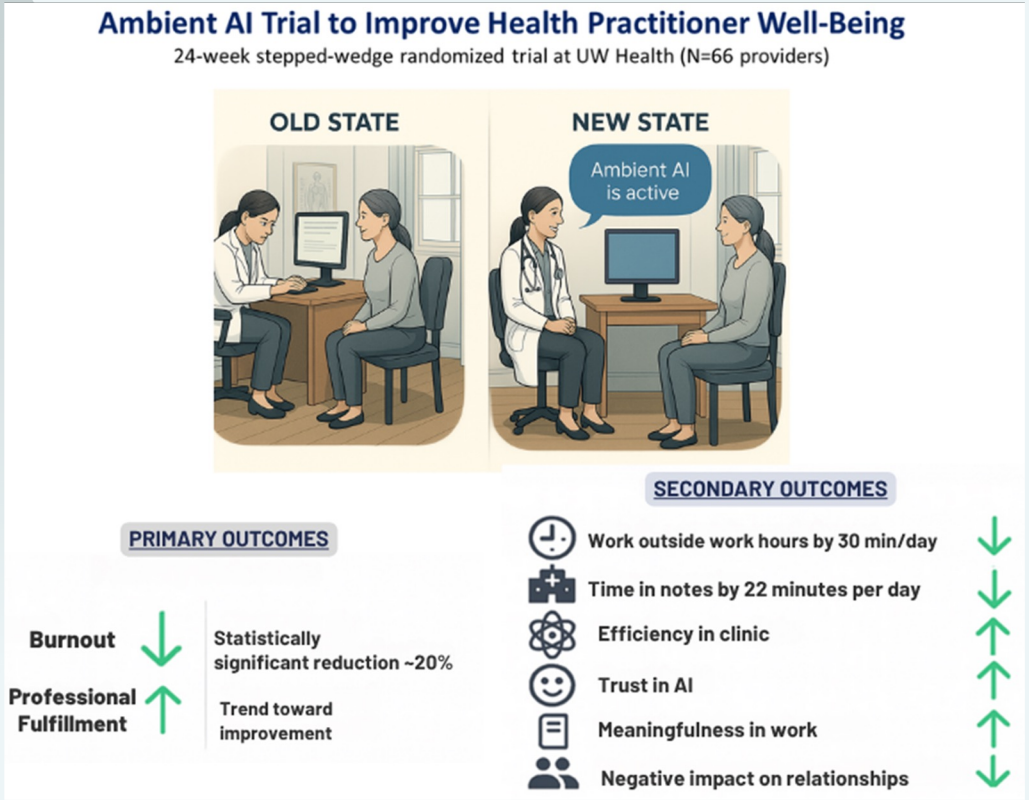


UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

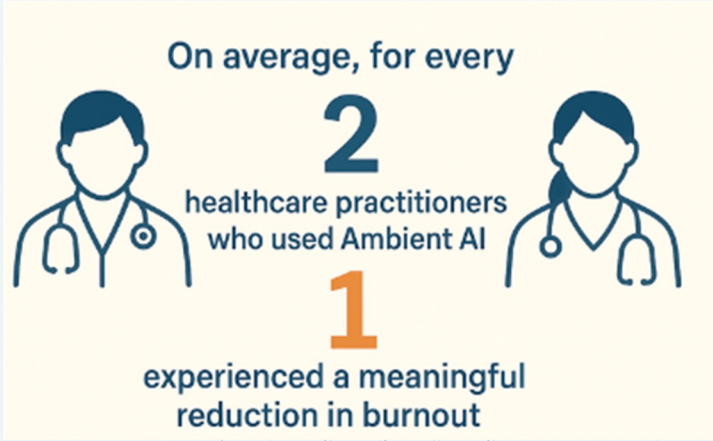
Secondary Outcomes: Relationships and Work Meaning



Trial Results Summary



Number need to treat



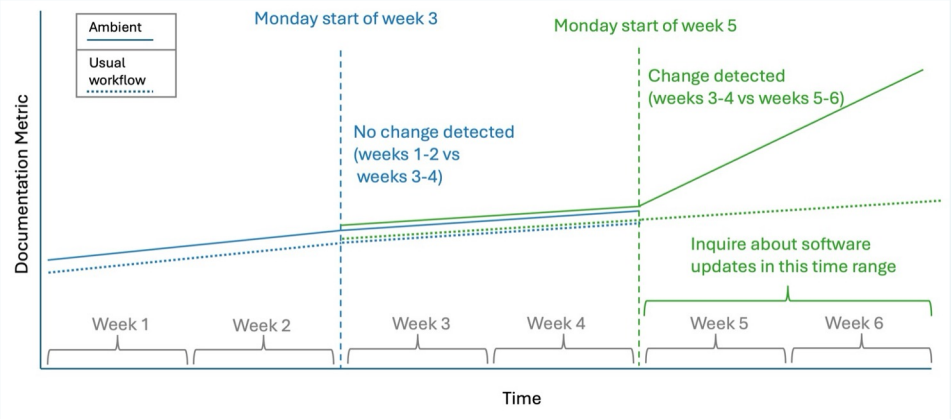
UW Health | School of Medicine and Public Health, University of Wisconsin–Madison

Additional Metrics Important to Operations for a LHS

UWHealth | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

Unplanned Event Monitoring that is from Health System perspective and agnostic to any AI software

- Ambient AI is a live, dynamic vendor-controlled tool
  - No guarantee that this (or any!) scribe in week 2 operates the same as scribe in week 15
  - No control or notice of sudden tool updates or bug fixes



- If you can't control it, (try to) track it
  - Monitor non-primary daily documentation/utilization metrics on rolling 2-week basis
  - Test for significant drifts via difference-in-differences modeling/testing framework
    - Positive flag triggers root cause analyses and discussions with operational team

Quality of Clinical Time Matters as Much as Quantity

- **Clinician Experience (Post-Trial Interviews)**
  - 15% of participants interviewed (n=10)
  - Software used most or nearly all encounters
  - Overall Experience: 7 very positive, 3 mixed but improving

- **How Time Was Used - Not Just How Much**
  - Greater focus on the patient
  - Lower cognitive load and stress
  - Higher quality clinical notes
  - WoW described as “rare” after adoption

- **Perceived Value to Professional Fulfillment**
  - Nearly all participants said they would be disappointed or consider leaving if the software was removed
  - Highlights impact on workflow satisfaction and meaning in work, not just efficiency

Improved ICD-10 Coding Quality with Ambient AI

- **Evaluation Design**
  - Stratified random sample of 6,110 notes
  - All 66 participants included
  - Balanced by practitioner and note type
  - Ambient use in 50% of sampled notes
  - Gold standard adjudication by professional coders
- **Primary Outcome: ICD-10 Compliance (0-10)**
  - Ambient AI-generated Notes
    - Mean score 6.87 (95% CI: 6.76-6.98)
  - Human Authored Notes:
    - Mean Score 5.94 (95% CI: 5.81-6.07)

- **Ambient was associated with higher alignment score ($p<0.001$) than humans between documentation and final billing diagnosis.**

Documentation Quality

- 7966 randomly sampled notes representative of all healthcare practitioners
- Average 6559 input tokens (SD 2436) and 1953 output tokens (SD 334).

- **Accuracy**
  - Fidelity of extraction and detection of any falsification or fabrication
  - Mean score was 4.44 (SD 0.93)

- **Thoroughness**
  - Assessing major omissions
  - Mean score was 4.57 (SD 0.68)

- **Abstraction/synthesis**
  - Ability to integrate and summarize across data elements
  - Mean score 3.97 (SD 0.58)

- **Usefulness**
  - Generation of information relevant and helpful for specialty-specific
  - Mean score was 4.83 (SD 0.51)

- **Linguistic quality**
  - Organization 4.90 (SD 0.16), Comprehensibility 4.99 (SD 0.13), and Succinctness 4.63 (SD 0.55)

- **Stigmatizing language** appeared in less than 1% of the notes (n=12).

Qualitative Findings of patient Experience

Methodological Consideration	**Description**
Setting	UW Health ambulatory clinics
Recruitment Approach	-MyChart messages sent to patients seeing a participating clinician in ongoing PCT -Random number generator used to determine sequence patient would be contacted about study -856 invitations sent, 54 patients expressed interest, 20 participated
Data Collection	-Zoom study visit -Survey (demographics, trust in AI) -Semi-structured interview -Data collection ends after meaning and code saturation
Data Analysis	-Deductive coding through Picker's patient experience framework and inductive coding

Courtesy of Oliver Nguyen, MSHI, ISyE PhD candidate

Patient Experience Qualitative Study

Key Themes Identified Across Interviews

1. **Respect for Patient Preferences**
    - Clinicians perceived as more present, attentive, and engaged
    - Reduced computer-mediated interaction enabled more open dialogue
2. **Information & Education**
    - Patients reported clearer explanations, more open-ended questioning, and deeper discussion of concerns
    - Enabled broader clinical conversations beyond the presenting complaint
3. **Continuity & Care Transitions**
    - Notes described as more readable, th
    - Improved understanding of care plan
4. **Contributors to Patient Comfort with AI:**

Themes	Examples
Patient-level demographic factors	- Having AI-related employment - Global openness to sharing information - Resemblance to familiar technologies
Saliency of AI scribe's physical presence	- Phone with AI scribe is often placed off to the side
Privacy-related concerns and beliefs	- Visit is not focused on sensitive topics - Minimal concerns on data privacy
Positive pre-existing relationships with clinicians	
Anticipated benefits	- Perceived improvement in clinical accuracy - Desire to improve clinicians' workflow

Real-time Data Monitoring in Clinical Practice

UW Health | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

Scalability and Realization in Clinical Practice

- What is our distribution?
  - Current: 556 faculty and APP providers
  - Target: 800

- What experiences have changed?
  - 208,447 notes created automatically
  - 64.2% of notes use ambient (of applicable notes)

- Margin Improvement – Financial Stewardship
  - Virtual Scribe usage is down from 250 to 35 providers and costs are down $4.4M
  - FY25/26 (and beyond) $2M budget reduction (funding through scribe savings)

UW and LHS Team:

NEJM AI

DATASETS, BENCHMARKS, AND PROTOCOLS

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

Authors: Majid Afshar, M.D., M.S.C.R. , Felice Resnik, Ph.D. , Mary Ryan Baumann, Ph.D. , Josie Hintzke, M.S. , Kayla Lemmon, M.S. , Anne Gravel Sullivan, Ph.D. , Tina Shah, M.D., M.P.H. , +17 , and Joel E. Gordon, M.D. Author Info & Affiliations

Published August 28, 2025 | NEJM AI 2025;2(9) | DOI: 10.1056/AIdbp2401267 | VOL. 2 NO. 9

ORIGINAL ARTICLE

A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being

Majid Afshar, M.D., M.S.,[1,2,3] Mary Ryan Baumann, Ph.D.,[1,4,5] Felice Resnik, Ph.D.,[1] Josie Hintzke, M.S.,[1] Anne Gravel Sullivan, Ph.D.,[1] Graham Wills, Ph.D.,[3] Kayla Lemmon, M.S.,[1] Jason Dambach, M.D.,[2,3] Leigh Ann Mrotek, Ph.D.,[1] Mariah Quinn, M.D., M.P.H.,[2,3] Kirsten Abramson, M.D.,[2,3] Peter Kleinschmidt, M.D.,[2,3] Thomas B. Brazelton, M.D., M.P.H.,[3,6] Margaret A. Leaf, M.S.,[3] Heidi Twedt, M.D.,[2,3] David Kunstman, M.D.,[3,7] Brian Patterson, M.D., M.P.H.,[3,8] Frank Liao, Ph.D.,[3,8] Stacy Rasmussen, B.S.,[3] Elizabeth S. Burnside, M.D., M.S.,[1,3] Cherodeep Goswami, M.B.A.,[3] and Joel Gordon, M.D.[3,7]

The Learning Loop

Institute for Clinical and Translational Research
UNIVERSITY OF WISCONSIN-MADISON

UWHealth | School of Medicine and Public Health
UNIVERSITY OF WISCONSIN-MADISON

NIH NCATS

https://ictr.wisc.edu/program/learning-health-systems-lhs/