Bennett Waxse, MD, PhD

Physician-Engineer | Clinical AI Agents, Evaluation, and Safety Systems
Board Certified: Internal Medicine, Pediatrics, Adult Infectious Diseases
bennettwaxse@gmail.com | Denver, CO | github.com/bwaxse | linkedin.com/in/waxse/

Professional Summary

Physician-engineer working on production clinical AI and the systems that measure it. At Doctronic I design and ship capabilities inside a multi-agent medical harness (diagnostic interview, prescription refill, emergency surveillance) and build the evaluation infrastructure that gates them. Producing a plausible clinical agent is cheap; knowing whether it works is not, and that's critical in medicine. Board-certified in internal medicine, pediatrics, and infectious diseases, with first-author work in computable phenotyping across 500M+ clinical observations. And soon, clinical AI evaluation.

Technical Skills

Agentic Systems: Multi-agent orchestrator/sub-agent architectures, Claude Agent SDK, Deep Agents harness, tool-use design, deterministic safety gates and completion contracts, context engineering, latency optimization, prompt design
Evaluation & Measurement: LLM-as-judge design and calibration, simulated-patient fidelity instruments, pre-registered study design, benchmark contamination detection, inter-rater reliability and paired significance testing for eval design; CRAFT-MD, HealthBench, MedQA-CS, AgentClinic, OSCE/PACES rubrics
Retrieval: BM25, hybrid dense+sparse with RRF, HyDE, agentic retrieval loops, DSPy, GEPA optimization, gold-set construction, R-precision / recall@k
Engineering & Clinical Data: Python (Polars, Streamlit), SQL, R, Docker, AWS (Bedrock, Fargate, S3), Azure AI Foundry, BigQuery, MCP servers, CI; OMOP CDM, SNOMED CT, LOINC, RxNorm, ATC, FDA NDC, FDB

Professional Experience

Senior Clinical Context Engineer March 2026 – Present
Doctronic, medical AI platform for consumers and health-system partners

Agentic clinical systems: design and implementation
- Co-designed the architecture of Doctronic's next-generation consultation system: an orchestrator coordinating specialized sub-agents, patient thread decoupled from agent thread, and safety enforced through deterministic contracts rather than prompt instruction. Reviewed with the CEO in weekly architecture check-ins
- Built the diagnostic interview capability, a live sub-agent emitting a ranked differential with supporting and refuting evidence plus next-best-question reasoning, along with the gate that keeps any diagnostic statement from reaching a patient without a produced differential
- Built the prescription refill flow mirroring the production state pipeline: renewal-safety interview, parallel safety engines over drug monographs, a monograph resolver vendoring production clinical data behind a drift guard, a structural red-flag gate, and a deterministic fill-versus-escalate decision. Split it into an independently shippable flow and am helping take it and the other consultation flows to production
- Designed and built the guardian sentinel: emergency surveillance split out of diagnostic reasoning into an always-on parallel track, so red-flag detection is not entangled with the working-theory differential. Cut consultation latency via a priority-ordered question queue (roughly 60–70% fewer model round-trips)

Evaluation infrastructure and measurement integrity
- Led external benchmark validation on CRAFT-MD, where the platform exceeded the base model for diagnostic accuracy on a multi-turn conversational benchmark
- Found a contamination in that benchmark before the result went out: simulated patients were disclosing their own artificiality to the diagnosing model, which inflated scores. Built a detection instrument, cut the leak by more than an order of magnitude, and made a three-axis patient-fidelity judge (meta-disclosure, hallucination, context failure) an upstream gate on every downstream result
- Rebuilt that comparison's design to survive review: one metric at two cutoffs applied identically to both arms, McNemar's exact test to match the paired design, a filter for vignettes whose answers were not diagnoses, and comparator configuration aligned with ours for a fair contest
- Found and fixed two defects silently corrupting results: per-case clinical data never reached the simulated patient's prompt, invalidating every prior run in a case family; and the routing signal had no schema validation, with enum conformance swinging 98% to 3% to 80% across three consecutive system versions
- Shipped the routing evaluation layer for multi-intent consultations: a dataset of conversations that pivot mid-dialogue, deterministic per-shape routing metrics, and a question-anchored clinical QA judge that finds where a patient asked and where the doctor substantively answered rather than assuming fixed turn positions
- Reviewed 208 cases from a widely cited clinical benchmark and concluded it was a poor fit for a conversational diagnostic product, since it penalizes information-seeking. That negative result redirected the external-validation strategy. Authored the specification that became the clinical evaluation framework, and validated LLM-as-judge against physician adjudication at ~88% agreement (about ten cents per conversation, cached)

Clinical knowledge retrieval and safety criteria

- Ran an evaluation of Doctronic's clinical guidance retrieval end to end, from problem framing to the recommendation the production team adopted, building a 300-case gold standard from real patient conversations via a two-stage nominate-then-verify judge with positive and negative controls and a human review gate
- Found that the best-performing architecture (generate a hypothetical guidance document in the corpus's own voice, retrieve lexically over it, then re-rank) reaches roughly 93% recall and over 80% precision at sub-second latency for a fraction of the full-catalog language-model cost. Traced where the advantage actually comes from rather than accepting the aggregate: query-side diversity and selection, not the agentic loop
- Wrote or substantially revised the clinical flag criteria for eight emergency presentations (acute coronary syndrome, aortic dissection, stroke, dyspnea, major trauma, acute psychosis, severe abdominal pain, self-harm risk), with boolean flag-combination logic and per-question exemplars. Helped diagnose a production screener over-firing defect generating over a thousand false positives in a month. Separately expanded the platform formulary from 189 to 3,177 medications (FDA NDC, RxNorm/ATC, FDB enrichment) with the reviewer application physicians used to adjudicate it

Team, hiring, and research program

- Technical interviewer and format owner for the Clinical Context Engineer roles; built the take-home from scratch: an 800-guideline synthetic clinical corpus with validated traps (omitted must-not-miss conditions, lexical decoys, near-neighbor pairs) and a grader-side check confirming naive retrievers collapse on the danger tail. Founded the organization's primary-source research corpus; contributed technical sections to health-system partnership and board presentations

Research Data Scientist & Clinical Informaticist August 2023 – March 2026
National Human Genome Research Institute, Precision Health Informatics Section, NIH

- Translated clinical knowledge into computable phenotype definitions used in the *All of Us* Research Program (500M+ observations), validating diagnostic accuracy at 79–97% PPV across 8 respiratory pathogens (first author, *Scientific Reports*, 2025)
- Built reproducible pipelines over large-scale real-world medication data, including outpatient utilization classification across roughly 1,000 ingredients. Coordinated 10+ clinician-scientist collaborations; mentored junior researchers on EHR methodology and terminology mapping across OMOP CDM, SNOMED CT, LOINC, and RxNorm

Infectious Diseases Clinical Fellow July 2021 – March 2026
National Institute of Allergy and Infectious Diseases, NIH

- Managed complex cases across urgent care, primary care, and inpatient settings spanning internal medicine, pediatrics, and infectious diseases, building the familiarity with diagnostic and prescribing logic that now grounds my agent and criteria design work

Medicine-Pediatrics Resident and Research Fellow June 2017 – June 2021
University of Chicago Medicine
Retrospective EHR analysis identifying differential treatment responses in septic shock (American Thoracic Society 2020); completed outcomes research training fellowship in biostatistics, epidemiology, and clinical research methods.

Education

Infectious Diseases Fellowship, NIAID, NIH 2021 – 2026
Medicine-Pediatrics Residency, University of Chicago 2017 – 2021
MD, UT Southwestern Medical School 2008 – 2017
PhD, Clinical Biochemistry, NIH-Cambridge Scholars Programme, University of Cambridge 2010 – 2016
BA, Biology and Chemistry, Texas Christian University, Honors 2004 – 2008

Selected Publications and Presentations

Pradhan A, **Waxse BJ**, Matias WR, *et al.* Evaluation of four large language models on complex, infectious disease case scenarios. *medRxiv.* 2026.

Fulda ES, **Waxse BJ**, Goleva SB, *et al.* 11 million days of longitudinal wearable data reveal novel future health insights. *medRxiv.* 2026.

Waxse BJ, Bustos Carillo FA, Tran TC, *et al.* Computable phenotypes to identify respiratory viral infections in the All of Us research program. *Scientific Reports.* 2025;15(1):18680.

Waxse BJ, Rao S. Data Science for Pediatric Infectious Disease: Utilizing COVID-19 as a Model. *Current Opinion in Inf Dis.* 2025.

Mo H, Channa Y, Ferrara TM, **Waxse BJ**, *et al.* Hyponatremia Associated with the Use of Common Antidepressants in the All of Us Research Program. *Clin Pharmacol Ther.* 2025;117(2):534-543.

Also co-author, *Clin Pharmacol Ther* (2024) and *medRxiv* (2026, variant association testing across 392,030 whole genomes); first author, *Journal of Cell Science*; co-author, *Nature Cell Biology.* Podium presentations: IDWeek 2026, AMIA 2026, IDWeek 2025, AMIA 2024.