David Hein

Dallas, TX | davidhein67@gmail.com
LinkedIn | GitHub | Portfolio

Professional Summary

Healthcare-focused applied AI/data scientist with 6 years of experience across clinical research data science and applied AI systems. First-author publications in npj Digital Medicine, JNCI, and Genome Medicine, with recent work spanning clinical NLP, agentic LLM workflows, data engineering, and applied AI applications. Hands-on across the path from messy raw data to production-facing application.

Experience

Data Scientist | UTSW Medical Center | September 2023 – Present

- Designed and published a novel human-in-the-loop error ontology for LLM-based clinical information extraction, enabling physicians and data scientists to identify failure modes, refine task definitions, and iteratively improve model outputs.
- Developed agentic LLM workflows for longitudinal oncology data that maintain auditable ledgers of treatment and disease events, structure evidence across dozens of records per patient, and support downstream progression-free survival analysis.
- Engineered reproducible Python pipelines for 200,000+ kidney cancer EMR records, converting raw Box-hosted exports into validated Parquet artifacts, DuckDB-backed datasets, and provenance-tracked inputs for AI workflows and clinical research.
- Built a high-throughput PHI de-identification pipeline around Presidio, spaCy, and Hugging Face transformers, adding ONNX support, multi-worker/multi-GPU processing, parquet streaming, date hashing/jittering, benchmark datasets, and fuzz tests for missed PHI patterns.
- Refactored Streamlit-based spatial biology workflow tools with background workers, cancellation support, parameter tracking, rehydration, log steaming, GPU-optimized long-running steps, and custom Svelte components for fast OME-TIFF thumbnail and overlay visualization.
- Co-architected and deployed a campus platform for research asset commercialization and investor discovery, implementing end-to-end on Azure with FastAPI, AzureSQL, Entra SSO, Application Insights, GitHub CI/CD, and AI assisted search features backed by RAG-style retrieval
- Teach practical AI workshops for clinicians, researchers, IR staff, and operational leaders, with emphasis on model failure modes, bias, evaluation methods, and responsible use of AI tools

Clinical Data Specialist | UTSW Medical Center | February 2020 – August 2023

- Served as the primary data scientist/research analyst for multiple oncology research projects, supporting study design and execution, statistical analyses, and bioinformatics workflows.
- Led cleaning, normalization, statistical inference, and visualization of targeted panel sequencing and whole-transcriptome RNA-seq data in a Tempus collaboration, contributing to a co-first-author Genome Medicine publication on genetic ancestry and molecular tumor profiles.
- Optimized large batch genomic processing workflows in HPC environments by tunning parallelization, memory allocation, and disk configuration, reducing runtime and platform costs by roughly 80%
- Managed regulated clinical research data under HIPAA, IRB, HSP, and GCP training requirements, including consent documentation, secure data handling, and compliant data transfers with external research partners.

Education

MS in Data Science, The University of Texas at Austin | May 2023

- Relevant coursework: deep learning, natural language processing, data structures & algorithms, advanced predictive modeling, probability, causal inference

BS in Chemistry, The University of Texas at Austin | December 2019

- High Honors; McCombs Business Foundations Certificate; department academic scholarship

Skills & Tech Stack

Languages: Python, SQL, Go, Bash, R, TypeScript

AI & NLP: LLM evaluation, clinical information extraction, agentic workflows, prompt engineering, RAG, tool/function calling, structured outputs, NER, document understanding, Hugging Face Transformers, spaCy, Presidio, Azure OpenAI, LangChain, LangGraph, MCP

Machine Learning & Model Serving: PyTorch, scikit-learn, deep learning, PEFT/LoRA, vLLM, model validation, error analysis, classification metrics, inference tuning, YOLO

Statistics: survival analysis, mixed-effects models, bootstrapping, multiple testing correction, regression analysis, hypothesis testing, clustering & unsupervised learning

Data Engineering & Lakehouse: Spark/PySpark, Apache Iceberg, Pandas, Polars, NumPy, SciPy, PyArrow, DuckDB, Parquet, dbt

Clinical / Research Data: Epic, i2b2, REDCap, FHIR/HL7, ICD-10, HIPAA, IRB protocols, GCP training, oncology real-world data, biospecimen-linked datasets

Backend, Databases & Storage: FastAPI, Flask, Pydantic, REST APIs, PostgreSQL, SQL Server, Azure SQL, SQLite, Redis, Azure Blob Storage, MinIO, Faiss, Chroma

Azure & Cloud Infrastructure: Azure (Azure OpenAI, Azure SQL, Azure Blob Storage, Azure Web App, Application Insights, Log Analytics, Entra ID, Bicep), Docker, Docker Compose, Linux, Ansible

MLOps & CI/CD: MLflow, Prefect, Git, GitHub Actions, GitLab CI, pytest, OpenTelemetry, model monitoring, drift detection, experiment tracking

Visualization & Reporting: Power BI, Plotly, Quarto, Jupyter, Streamlit, Dash, ggplot2

Selected Publications

1. **Hein D**, Christie A, *et al*. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *npj Digital Medicine*. 2025.
2. Jamieson AR, Holcomb MJ, …, **Hein D**, *et al*. Rubrics to prompts: assessing medical student post-encounter notes with AI. *NEJM AI*. 2024.
3. Rhead B, **Hein D***, *et al*. Association of genetic ancestry with molecular tumor profiles in colorectal cancer. *Genome Medicine*. 2024. (*co-first author)
4. Sanford NN, Shi Q, **Hein D**, Hall WA. Benchmarks of success in radiotherapy vs systemic therapy: National Clinical Trials Network randomized controlled trials sponsored by the National Cancer Institute. *Journal of the National Cancer Institute*. 2025.
5. **Hein D**, Coughlin LA, *et al.* Assessment of distinct gut microbiome signatures in a diverse cohort of patients undergoing definitive treatment for rectal cancer. *Journal of Immunotherapy and Precision Oncology*. 2024.

Full publication record: ORCID

Selected Honors, Open Source, & Service

- Awarded $15,000 from the UTSW Office of Technology Development as team lead for NIH I-Corps customer discovery and commercialization planning for an AI medical education platform.
- Built InstaWell, an open-source Python/Dash/Plotly toolkit for analyzing thermal shift assay data.
- Designed and maintain the WordPress website for Feeding Families of Alabama, a regional food bank and nonprofit.