Xiyu Ding

dxy_03@outlook.com | (857) 999-5736 | https://www.linkedin.com/in/grace-xiyu-ding/
520 Park Ave, Baltimore, MD 21201

PROFILE

- PhD researcher with 7 years of experience developing and applying AI/ML techniques to problems in the clinical and biomedical sciences
- Proficient in Git, HPC, Linux, R and Python (pytorch, matplotlib, pandas, numpy, scikit-learn, etc.)
- Research topics: **Causal Inference, Hierarchical Bayesian Models, Natural Language Processing, Multimodal Modeling, Real World Evidence, Federated Transfer Learning**

EDUCATION

Johns Hopkins University, Biomedical Informatics and Data Science **Baltimore, MD**
PhD in Health Informatics, GPA: 3.99 *Sep 2021- Aug 2026 (Expected)*

Harvard University, Department of Biostatistics **Boston, MA**
MS in Computational Biology, GPA: 3.89 *Sep 2018 - May 2020*

Nanjing University School of Life Science **Nanjing, China**
BS in Biological Science, GPA: 3.92 *Sep 2014 - Jun 2018*

EXPERIENCE

Johns Hopkins University **Baltimore, MA**
Graduate Researcher (AI / ML / Health Data) *Sep 2021 – Present*

- Developing a **scalable sparse representation learning pipeline** to cluster patients from high-dimensional binary Dx/Rx data (**PCA, NMF, K-medoids**) and integrate cluster features into healthcare utilization prediction.
- Developed a **Bayesian hierarchical transfer learning framework** with global-local skew-shrinkage priors to improve prediction & inference in small/under-represented populations, achieving >**50% performance gains in clinical and genetic studies**
- Led **federated analytics** across 10+ international healthcare databases (OHDSI/LEGEND-T2DM; >5M patients) to evaluate sex differences in the **comparative effectiveness and safety** of second-line T2DM therapies
- Built **multimodal deep learning models** integrating 3D MRI, deformation (Jacobian) maps, and biomarkers via cross-attention for Alzheimer's disease classification, outperforming transformer baselines with fewer parameters.
- Benchmarked imaging preprocessing pipelines and architectures (CNN, ResNet, ViT) under limited data/compute settings to guide model selection
- Designed interpretability workflows (Grad-CAM++, attention visualization) to support model validation and clinical insight

Eli Lilly and Company **Indianapolis, IN**
Research Intern, Advanced Intelligence *May 2025-Aug 2025*

- Explored and benchmarked time-series forecasting models (**ETS, Theta, iETS, XGBoost, Random Forest**) for **intermittent and sparse demand**, addressing volatility and regime shifts
- Designed **simulation and stress-testing frameworks** to evaluate forecasting model robustness under spikes, drops, and distributional shifts, informing risk-aware forecasting strategy
- Built reproducible Python/R pipelines to support large-scale forecasting experiments and rapid model comparison across product portfolios
- Partnered with supply-chain and operations stakeholders to translate forecasts into actionable planning insights, balancing accuracy, interpretability, and operational constraints

Computational Health Informatics Program, Boston Children's Hospital **Boston, MA**
Data Scientist *Jul 2020 - Aug 2021*
Work/Study Intern *May 2019 - May 2020*

- Designed and deployed a **transformer-based model** to identify adult congenital heart disease diagnoses and phenotypes from long, unstructured EHR narratives
- Applied **Longformer and hierarchical attention architectures** to handle long clinical document classification, improving accuracy by 20% and interpretability over traditional coding approaches
- Fine-tuned **RoBERTa-based models** for psychiatric sentiment extraction, incorporating domain-specific risk factors and achieving a 17% improvement in macro-F1
- Developed **multi-task BERT-based models** for automated triage and referral attribute classification, improving classification macro F1 score from 0.63 to 0.81; Explored prompt-based sequence-to-sequence models as an alternative paradigm, comparing prompt-driven classification against fine-tuned transformers

Department of Data Science, Dana-Farber Cancer Institute **Boston, MA**

Graduate Research Assistant/Statistical Programmer *May 2019 - May 2020*

- Refactored R package to compute cancer risk associated with germline mutations for ASK2ME™
- Implemented an end-to-end R data processing pipeline to increase computational efficiency by 80%

INDEPENDENT /APPLIED AI PROJECTS

Weiyi LLC **Shanghai, China**

Co-Founder *2025 Sep - Present*

- Built **agentic LLM systems** for automated survey data analysis, including data cleaning, visualization, clustering and phenotyping, predictive modeling and survival analyses (collaboration with China CDC, Fudan University)
- Developed **LLM + RAG pipelines** for automatic ICD-10 code extraction from clinical narratives, combining rule-guided retrieval with evidence-grounded reasoning; designed **multi-step agent workflows** (evidence extraction → candidate generation → rule validation) to improve precision, auditability, and clinical alignment (collaboration with Henan Cancer Hospital)

SELECTED PUBLICATIONS

1. S Zhang, **X Ding**, K Ding, J Zhang, K Galinsky, M Wang, RP Mayers, Z Wang, H Kharrazi. ProtoBERT-LoRA: Parameter-Efficient Prototypical Finetuning for Immunotherapy Study Identification. **AMIA annual symposium 2025, Atlanta, GA. [Oral Presentation, Distinguished Paper Award]**
2. S Zhang, **X Ding**, B Caffo, J Chen, C Zhang, Z Wang, H Kharrazi. Cross-Attention Fusion of MRI and Jacobian Maps for Alzheimer's Disease Diagnosis. arXiv preprint arXiv:2503.00586, 2025. **[Accepted at ISBI 2026]**
3. **Ding X**, Kharrazi H, Nishimura A. Assessing the impact of social determinants of health on diabetes severity and management. **JAMIA Open.** 2024 Oct 25;7(4):ooae107.
4. **Ding X,** Barnett M, Mehrotra A, Tuot DS, Bitterman DS, Miller TA. Classifying unstructured electronic consult messages to understand primary care physician specialty information needs. **J Am Med Inform Assoc.** 2022 Aug 16;29(9):1607-1617. doi: 10.1093/jamia/ocac092.
5. Su X, Miller T, **Ding X**, Afshar M, Dligach D. Classifying long clinical documents with pre-trained transformers. arXiv preprint arXiv:2105.06752, 2021
6. **Ding X**, Hall MH, Miller T. Incorporating Risk Factor Embeddings in Pre-trained Transformers Improves Sentiment Prediction in Psychiatric Discharge Summaries. **In Proceedings of the 3rd Clinical Natural Language Processing Workshop (pp. 35–40). ACL. 2020.** [Oral Presentation]
7. **Ding X**, Barnett M, Mehrotra A, Miller T. Methods for Extracting Information from Messages from Primary Care Providers to Specialists. **In Proceedings of the 1st Workshop on Natural Language Processing for Medical Conversations (pp. 1–6). ACL. 2020.** [Oral Presentation]
8. Liu D, Clemente L, Poirier C, **Ding X**, Chinazzi M, Davis J, Vespignani A, Santillana M. Real-Time Forecasting of the COVID-19 Outbreak in Chinese Provinces: Machine Learning Approach Using Novel Digital Data and Estimates from Mechanistic Models. **J Med Internet Res.** 2020 22(8), e20285.

MANUCRIPT in PREPARATION

1. **Ding X**, Gu Y, Chin A, Nishimura A. High-dimensional Bayesian transfer learning through skew shrinkage priors: applications to genetic and electronic health records data. Annal of Applied Statistics.
2. **Ding X**, et al. Sex Differences in the Comparative Effectiveness and Safety of Second-line Anti-diabetic Agents: Real-world Evidence from Large-scale Multinational Study. JAMA Internal Medicine.
3. Liu S, **Ding X**, Lehmann HP. Bootstrapping Bezier Curves to Quantify the Uncertainty Associated with the Decision-Analytic Optimal Point on the Receiver Operating Characteristic Curve. Medical Decision Making.
4. Chin A, **Ding X**, Nishimura A. Spectral Collapsed Gibbs Sampler for Global-Local Shrinkage Priors.

PROFESSIONAL SOCIETY MEMBERSHIP

Association for Computational Linguistics (2020-2021)
American Medical Informatics Association (2020-2023, 2025-2026)

TEACHING EXPERIENCE

Natural Language Processing in the Health Sciences (Johns Hopkins University, Spring 22, 24)
Clinical Data Analysis with Python (Johns Hopkins University, Fall 22, 23, 24)
Professional Certificate in Data Science, Data Analysis for Life Science Series (Harvard X, May 2019-Aug 2019)

SELECTED AWARDS

✧ 2025 Global Biotech Revolution "Leader of Tomorrow" (1 of 100 selected globally, 2025)
✧ Distinguished Poster Nomination at AMIA 2020 Virtual Annual Symposium (Nov 2020)
✧ Hackathon Track Winner (Artificial Intelligence in Healthcare) and Health Equity and Leadership Hackathon Award at 2019 Hacking Public Health, Harvard T.H. Chan School of Public Health (2019)