Lessons Learned in Federated Cancer Research Pilots Spanning Four NIH-Designated Comprehensive Cancer Centers

Sahil Nalawade, James Wen, Alexander Chowdhury, Renato Umeton, Brian Wolpin, Michael Rosenthal
Dana-Farber Cancer Institute, Boston, MA

Introduction:

Federated learning (FL) presents a promising avenue for advancing healthcare data analysis, facilitating model training, inference, and collaboration across distributed datasets while upholding institutional data boundaries. FL decentralizes the learning process, ensuring data security and confidentiality, ideal for medical research and diagnosis where data privacy and limited data availability are paramount concerns. This abstract explores FL applications in healthcare data analysis, emphasizing model **inference**, **federated learning**, and **collaboration**.

Applications:

FL methods offer advantages in healthcare data analytics, categorized into federated model inference, federated model training, and collaborative FL utilization with external partners. In the contemporary AI landscape, models' complexity necessitates testing on diverse data, including multi-institutional datasets, posing challenges with IRB regulations and data sharing agreements. Federated model **inference** mitigates these challenges, enabling model evaluation across institutions while keeping data localized at each site. Our collaborative FL projects among NIH-designated cancer centers leverage Rhino Health's federated learning platform for various workflows, including data processing, DICOM annotation, inference, and training. We have conducted pilots for federated inference using fully-trained models to assess performance on multi-institutional data, addressing tasks like body composition analysis and multi-organ segmentation.

Federated learning uses data from multiple independent repositories to securely train global models. Our team is collaborating with UCSF to jointly train a clinical risk prediction model based on Large Language Model analysis of local clinical notes. This approach maintains patient privacy by training on clinical data from multiple institutions while centralizing the shared model updates, potentially enhancing the generalizability of the integrated model. Federated learning's distributed nature holds potential for enhancing clinical risk prediction while preserving privacy.

Federated learning facilitates **collaboration** between institutions by lowering the privacy risks and decreasing the administrative burdens associated with bringing new sites into a network. In our partnership with Rhino Health, the one-time investments in information security and data privacy reviews at each site have allowed for rapid iteration on new lines of research with those partners. Applications include testing pre-trained models on multi-institutional datasets for tasks like body composition analysis.

Conclusion:

Federated methods in healthcare data analysis show promise for accelerating the pace and generalizability of AI research while upholding patient privacy and site control of data. We demonstrate three applications in federated inference, model training, and collaboration, effectively executed using a federated learning framework.