Count Everything: Secure Count Query Framework Across Big Data Centers

Ida Sim, University of California, San Francisco*; Xiaoqian Jiang, University of California, San Diego; Claudiu Farcas, University of California, San Diego; Kevin Osborn, University of California, Santa Cruz; David Steinberg, University of California, Santa Cruz; Benedict Paten, University of California, Santa Cruz; Mark Diekhans, University of California, Santa Cruz; Paul Avillach, Harvard University; Wallace Wadge, Overlap Health; David Haussler, University of California, Santa Cruz; Isaac Kohane, Harvard University; Lucila Ohno-Machado, University of California, San Diego*

Electronic health record (EHR), mobile health and genome sequence data for the same patient are often spread across multiple geographically distributed hosts that expose different application program interfaces (API). The lack of secure and privacy-preserving interfaces impedes the ability of applications to integrate heterogeneous data from different sources. Because all of statistics start with counts of observations, developing a federated count query infrastructure is a first step to supporting more advanced queries.

The Count Everything project supports count query interoperability across three BD2K Centers [1]: the Mobile Sensor Data to Knowledge (MD2K) Center houses mobile health data [2]; the Center for Big Data in Translational Genomics (CBDTG) [3] houses genomic sequence data; the Patient-centered Information Commons: Standardized Unification of Research Elements (PIC-SURE) Center [4] houses EHR data. The NIH National Center for Biomedical Computing Integrating Data for Analysis, Anonymization, and Sharing (iDASH) [5] and the BD2K Biomedical and healthcare Data Discovery Index Ecosystem (bioCADDIE) [6] -- also participate in this project. We developed a secure protocol to decompose a user query into subqueries and return the cardinality of matched items that share the same standardized patient IDs. The three data-holding BD2K Centers exposed data to the Count Everything infrastructure via their respective APIs.

We created a synthetic cohort by assigning unique global IDs to convenience sets of mobile, genomic 1000 Genomes, and clinical CDC NHANES data, which are queried by the Count Everything infrastructure. To safeguard the intermediary results and return only the final count, we developed a secure framework based on homomorphic encryption (HME) [7], which can support computation on encrypted data. This is a semi-centralized system, which utilizes iDASH as the HME processor and bioCADDIE as a cryptographic service provider (CSP) in a hub-and-spoke architecture. The web-based service is available for testing: http://counteverything.ucsd-dbmi.org:8080/JavaBridge/beta2/index.html#.