MedCo^2: Privacy-Preserving Cohort Exploration and Analysis

David FROELICHER [a], Mickaël MISBACH [a], Juan R. TRONCOSO-PASTORIZA [a], Jean Louis RAISARO [b], and Jean-Pierre HUBAUX [a]

[a] *EPFL, Switzerland*
[b] *CHUV, Switzerland*

Abstract. Medical studies are usually time consuming, cumbersome and extremely costly to perform, and for exploratory research, their results are also difficult to predict a priori. This is particularly the case for rare diseases, for which finding enough patients is difficult and usually requires an international-scale research. In this case, the process can be even more difficult due to the heterogeneity of data-protection regulations, making the data sharing process particularly hard.

In this short paper, we propose MedCo^2 (pronounced *MedCo square*), a distributed system that streamlines the process of a medical study by bridging and enabling both data discovery and data analysis among multiple databases, while protecting data confidentiality and patients' privacy. MedCo^2 relies on interactive protocols, homomorphic encryption and differential privacy. It enables the privacy-preserving computations of multiple statistics such as cosine similarity and variance, and the training of machine learning models, on patients that are obliviously selected according to specific criteria among multiple databases.

Keywords. data discovery, data analysis, privacy, confidentiality, homomorphic encryption

1. Introduction

The current trend towards personalized medicine requires researchers to access increasingly larger cohorts of subjects [1], the size of which often exceeds the amount of data available at the researchers' facility. This results in an urgent need for efficient and automatic data sharing mechanisms between medical institutions. Medical data, however, are particularly sensitive, and attempting to share them poses significant threats to the subjects' privacy, as it increases the risk of leaks to ill-intentioned persons or organizations. Medical institutions often refrain from sharing their data due to the increasing number of breaches [2,3], and to their derived reputation and financial implications [4]. Finally, national and international regulations (e.g., HIPAA [5] in the U.S. and the GDPR [6] in the E.U.) impose strong requirements in terms of both confidentiality and restriction of data access that are commonly not met by the existing sharing-solutions [7]. In this environment, in order to access the data, researchers are faced with complex challenges throughout the whole process. Cohort exploration is usually the first step of a medical study; when performed across several medical institutions, it becomes very difficult, as obtaining the necessary legal authorizations and signing agreements between institutions

can take months. This becomes almost impossible when hospitals are in different countries because of the incompatibility of national legal frameworks. Finally, researchers can be left with data of limited value to perform the desired analyses, as they cannot get a priori information (e.g., size, availability and statistical properties such as distribution or moments) to assess the suitability of the data before actually gaining access to them. We propose MedCo^2, a novel system featuring privacy-preserving cohort exploration and advanced cohort analysis. It enables a researcher to assess the presence of subjects with specific traits and features, compute statistics, and/or train and evaluate machine-learning models on subjects' records, over distributed databases. By relying on distributed and interactive protocols, homomorphic encryption and differential privacy, our system ensures that these queries can be executed yet still protect the confidentiality of data and the privacy of subjects. Finally, MedCo^2 distributes the trust and avoids single points of failure, as both the storage and the computations are distributed among multiple nodes.

2. System and Threat Models

MedCo^2 supports a network of mutually distrustful medical institutions that act as data providers (DPs) and hold subjects' records. An authorized researcher (see Figure 1) can run queries without threatening the data confidentiality and subjects' privacy. The DPs are considered semi-honest and the researcher is considered malicious. If required, our system can also use zero-knowledge proofs and an immutable distributed ledger to cope with malicious DPs.

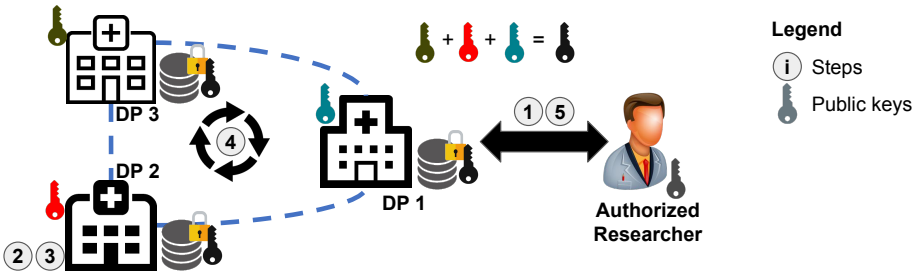

Figure 1. System Model.

3. Proposed Solution

The two main building blocks in MedCo^2 are MedCo [8] and Drynx [9]. MedCo is an operational distributed system that enables authorized researchers to explore cohorts distributed across several institutions, by filtering subjects with inclusion/exclusion of clinical and genetic criteria. Drynx enables the computation of statistics, such as standard deviation or extrema, and the training and evaluation of machine-learning models on distributed data. In order to ensure data confidentiality and individuals' privacy, both systems rely on distributed interactive protocols, homomorphic encryption and differential privacy. In MedCo^2, we combine MedCo's capabilities to privately identify a set of sub-

jects matching precise criteria with Drynx's ability to perform computations on the identified subjects' records. As a result, we provide a fully-decentralized system that enables multiple functionalities and enforces privacy and security guarantees that are stronger than in any (to the best of our knowledge) existing solutions [10,11].

We now briefly describe the MedCo^2 data model, before explaining how a query is executed. DPs have two separate databases: (1) the local database, that is in cleartext and is only accessible from within the institution, and (2) a research database, open to external queries, in which data are encrypted. We define a micro-ETL (extract, transform, load) process that encrypts selected data from the local database and transfers them to the research database. In our system, the data stored in the research databases and all those involved in the query execution are homomorphically encrypted under the DPs' collective key, as described in [9]. A researcher creates a query containing the subjects' selection criteria along with the computation she wishes to perform on the selected subjects' records (step 1 in Figure 1). This query is broadcast to the DPs who privately retrieve the list of subjects satisfying the query and initiate a local data retrieval process (step 2). In this process, each DP stacks the received query and the retrieved subjects' identifiers inside its research database. This stack is periodically emptied by the local operating database that pulls the waiting queries and authorizes them (e.g., based on regulations, local policies). If the query is accepted, each DP performs some query-dependent precomputation on its local database, before encoding and encrypting its contribution to the secure protocol (step 3). These contributions are then collectively aggregated under homomorphic encryption, to obtain the encrypted results (step 4). At this point, the results are obliviously switched from the collective key to the researcher's private key (Collective Key Switching protocol), and sent back to her (step 5). MedCo^2 executes, in parallel to the query execution, a micro-ETL in which it encrypts the attributes involved in the query and stores them in the research database. This means that when a DP receives another query on the same attributes, it can directly answer from the research database, thus reducing its workload for upcoming queries.

4. Evaluation

We implement and showcase MedCo^2 within a representative scenario that involves clinical and genetic data sharing. We integrate it in MedCo by enhancing its data model and query language/interpreter to seamlessly enable an authorized researcher to execute operations on a cohort built from MedCo.

We present MedCo^2 performance for a worst-case scenario: a query that it has not seen before, i.e., we assume that all the computation happens *on the fly*, with no precomputation. Figure 2 shows the timeline of a query execution in MedCo^2, comprising the steps of (a) query creation at the client, (b) query broadcast to the DPs, (c) collective query reencryption at the DPs (in order to enable matching queries [8]), (d) local DB execution and result retrieval at each DP, (e) local encoding of the matching patients' data at each DP, (f) collective aggregation of the target function across all DPs, (g) collective key switching of the results to the querier's key, and (h) decrypting and decoding at the querier. In this case, 18132 patients satisfy the filtering criteria and are retrieved among 150,000 patients distributively stored in 12 databases. As shown in [8,13], the overhead time for data discovery (including Step 1, which comprises Query Creation,

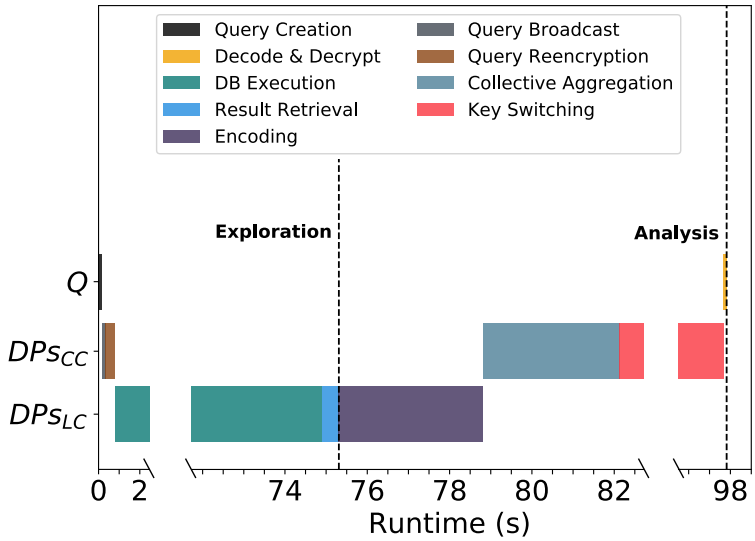

Figure 2. MedCo^2 Query Execution Timeline. *LC* =Local Computations, *CC* =Collective Computations.

Query Broadcast and Query Reencryption, and Step 2, composed of Database Execution and Result Retrieval) when executed on encrypted data is negligible, in this case 2 seconds, with respect to the same operation executed on cleartext data. Figure 2 also shows that MedCo^2 can perform a logistic regression on a distributed dataset with 100 features in less than 23 seconds. The distributively trained models achieve accuracy results similar to their centralized, non-secure counterparts, as shown in [9]. We observe that the whole MedCo^2 process is performed in less than 98 seconds if 18132 patients out of 150000 satisfy the selection criteria. The *DPs*' local query execution in MedCo (DB execution in Figure 2) is the most expensive operation, as it linearly depends on the number of selection criterias in the query and on the total number of observations in the database (each patient has thousands of observations).

Finally, MedCo^2 enables a researcher to train and evaluate linear and logistic regression models on a chosen cohort without accessing the actual subjects' records. This makes the task more cost-effective and streamlines a process that is at the core of many common medical workflows such as GWAS.

5. Conclusion

MedCo^2 addresses the privacy and security challenges of highly distributed medical-data sharing networks by avoiding single points of failure and by providing the following unparalleled functionalities and properties: (a) privacy-preserving cohort exploration and analysis over distributed databases held by distrustful DPs, (b) distributed training of simple machine learning models, (c) data confidentiality and individuals' privacy through collective keys and distributed protocols, and (d) a usable example scenario through its integration in a deployed system featuring a modern graphical user interface [8]. MedCo^2 aims at enabling the promises of personalized health by facilitating and streamlining the data discovery and analysis processes.

Acknowledgements

This work was supported in part by the grant #2017-201 (DPPH) of the Swiss strategic focus area Personalized Health and Related Technologies (PHRT), and by the grant #2018-522 (MedCo) of the PHRT and the Swiss Personalized Health Network (SPHN).

References

[1] P. Mahon and J.M. Tenenbaum, Paths to Precision Medicine - A Perspective., *Journal of Precision Medicine*, 2015.

[2] Ponemon Institute. Sixth annual benchmark study on privacy security of healthcare data, *Tech. report*, 2016.

[3] Healthcare Data Breach Statistics, *https://www.hipaajournal.com/healthcare-data-breach-statistics/*, Accessed: July 03, 2019.

[4] A huge trove of medical records [...] found exposed, *https://tcrn.ch/2Cp6hzv*, Accessed: July 15, 2019.

[5] The health insurance portability and accountability act (hipaa), *https://www.hhs.gov/hipaa/index.html*, Accessed: June 26, 2018.

[6] EU Parlament. The EU General Data Protection Regulation (GDPR). *http://www.eugdpr.org/*. Accessed: July 15, 2019.

[7] SHRINE, *https://www.i2b2.org/work/shrine.html*, Accessed: June 29, 2018.

[8] J.L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. A. Sá Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. A. Ford, and J.-P. Hubaux, Medco: Enabling privacy-conscious exploration of distributed clinical and genomic data. *IEEE/ACM TCBB*, 2018.

[9] D. Froelicher, J. R. Troncoso-Pastoriza, J. A. Gomes de Sá E Sousa and J.-P. Hubaux, Drynx: Decentralized, Secure, Verifiable System for Statistical Queries and Machine Learning on Distributed Datasets. *http://arxiv.org/abs/1902.03785*, 2018.

[10] H. Cho, D. J. Wu and B. Berger. Secure genome-wide association analysis using multiparty computation, *Nature biotechnology*, 2018.

[11] K. A. Jagadeesh, D. J. Wu, J. A. Birgmeier, D. Boneh and G. Bejerano. Deriving genomic diagnoses without revealing patient genomes. *Science*, 357(6352), 692-695, 2017.

[12] i2b2, *https://www.i2b2.org/*, Accessed: June 29, 2018.

[13] D. Grishin, J. L. Raisaro, J. R. Troncoso-Pastoriza, K. Obbad, K. Quinn, M. Misbach, J. Gollhardt, J. Sa, J. Fellay, G. M. Church and J.-P. Hubaux, Citizen-Centered, Auditable, and Privacy-Preserving Population Genomics, *bioRxiv, https://www.biorxiv.org/content/10.1101/799999v1*.