HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2020 January 10.

Published in final edited form as:

Pac Symp Biocomput. 2020 ; 25: 647–658.

Implementing a Cloud Based Method for Protected Clinical Trial Data Sharing

Gaurav Luthria[*], **Qingbo Wang**[*]

Department of Biomedical Informatics, Harvard University, Cambridge, MA 02138, USA

Abstract

Clinical trials generate a large amount of data that have been underutilized due to obstacles that prevent data sharing including risking patient privacy, data misrepresentation, and invalid secondary analyses. In order to address these obstacles, we developed a novel data sharing method which ensures patient privacy while also protecting the interests of clinical trial investigators. Our flexible and robust approach involves two components: (1) an advanced cloud-based querying language that allows users to test hypotheses without direct access to the real clinical trial data and (2) corresponding synthetic data for the query of interest that allows for exploratory research and model development. Both components can be modified by the clinical trial investigator depending on factors such as the type of trial or number of patients enrolled. To test the effectiveness of our system, we first implement a simple and robust permutation based synthetic data generator. We then use the synthetic data generator coupled with our querying language to identify significant relationships among variables in a realistic clinical trial dataset.

Keywords

Synthetic data; data sharing; patient privacy; clinical trials; cloud computing

1. Introduction

Clinical trials are used to evaluate the safety and efficacy of new medical technologies or treatments. Despite the fundamental role that clinical trials play in advancing medicine, access to a majority of clinical trial data has been restricted. Data sharing can improve clinical care as well as lead to new developments in administering clinical trials, personalized treatment strategies, and improved modifications to the technologies or treatments currently being evaluated. In addition to increasing scientific knowledge, sharing clinical data can improve the timeliness of data analysis, spark new research ideas, and decrease expenditures by avoiding unnecessary duplicate trials.1,2 Clinical trials produce a large amount of data including patient demographics, lab reports, and drug exposure, which are currently being underutilized due to the difficulty in ensuring patient privacy.3,4

gluthria@g.harvard.edu.

[*]These authors contributed equally to this work.

In addition to the primary concern of protecting patient privacy, some secondary concerns of data sharing include other investigators taking false ownership of clinical trial discoveries or invalid secondary analyses.5,6 The current solution for clinical trial data sharing that addresses the concerns mentioned above often involves detailed data request proposals and lengthy proposal approval times, hindering exploratory data analysis.7 To date, there has been no well known implemented computational system that addresses these concerns while providing public access to clinical trial data.

There are currently two mechanisms in place that allow for protecting patient privacy during data sharing: (1) explicit patient consent and (2) de-identification. Patient consent for post-hoc research studies is often impractical because it involves tracking a large number of patients that may result in bias between consenter and non-consenter populations.8 De-identification involves perturbing the data via statistical models or by removing identifying information.9–11 For example, electronic health records (EHR) in the United States are de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA). HIPAA requires 18 different fields representing “unique identifying characteristics” such as patient name, date of birth, and date of visit be removed prior to data sharing. Despite the extraction of such fields, patients are still susceptible to re-identification attacks.12 Protecting patient privacy is essential for maintaining trust between patients and healthcare professionals as well as preventing potential stigmatization or discrimination based on patient health.13

Automated de-identification models can be subdivided into three categories: (1) rule based models (2) machine learning models, and (3) hybrid models.14 Rule based models involve applying a set of curated rules, often developed by healthcare professionals, to perturb the real clinical data. Machine learning models apply probabilistic or off-the-shelf machine learning methods such as markov random fields, support vector machines, decision trees, and regression based models.15 Recently, deep learning has also been used to generate synthetic datasets.16 Hybrid models use a combination of rule based and machine learning models. Despite recent success in generating synthetic datasets, each method has inherent disadvantages. They may be application specific, dependent on curation or verification by medical professionals, difficult to implement, or produce inconsistent results compared to the real data in downstream analysis. These problems are further exacerbated when dealing with data from patients with rare diseases, as clinical data on these patients is much more specific and limited.

Simulating real-life clinical data is an extremely diffcult, if not an impossible task. To solve this inherent problem of balancing data sharing while ensuring that patient privacy is protected, we propose a platform where researchers can integrate features from both synthetically generated and real clinical datasets to conduct both exploratory and hypothesis driven research. Furthermore, we aim to address additional non-privacy related concerns of data sharing by developing a flexible, compartmentalized system that can be modified by clinical trial investigators. The specific contributions of our work are as follows:

- an integratable multicomponent system for researchers to access advanced statistics for real patient data while also providing the corresponding synthetic data for the query of interest
- a querying language for obtaining advanced summary statistics from the real patient dataset
- a simple, robust, and easily implementable permutation based algorithm for synthetic data generation

We first describe how each of these components are developed and how they can be integrated together while maintaining patient privacy and providing the most power to researchers. We next evaluate our system's querying language and compare the generated synthetic data to the real data. We finally demonstrate how our system can be used for research. We envision that the following system will enable public online sharing of clinical data without restricted access.

2. Methods

2.1. Storing clinical trial data in the cloud

Our system does not have a centralized storage unit. Clinical trial investigators can store their data in their own bucket without risking privacy violations. In other words, there is no need to upload the data to a certain platform. Instead, investigators only need to change the read-access of their clinical trial data so our execution system can access this data.

Though access to the real clinical trial data is restricted, we provide a highly flexible query language that enables all users to submit a job from a personal computer that runs on the execution system in the cloud (Figure 1A). Using this query language, users can investigate clinical data submitted by different investigators in a flexible way, such as performing statistical tests or visualizing distributions. Summary level data is also downloadable. The risk of privacy-sensitive information being identified or original data being reconstructed from the combination of summary data is minimized by restricting the type of queries users can make and by setting a minimum threshold for the query to be valid. For example, the maximum and minimum of a particular feature in the data can only be obtained when the sample size is greater than five individuals. Additionally, our query language allows users to download the corresponding synthetic data that closely mimics the original clinical trial dataset (Figure 1B). In methods 2 and 3, respectively, we will further describe our query language and how synthetic data was generated.

2.2 Flexible query language for advanced statistical analysis

Our query language uses Hail (https://hail.is), an open-source, Python-based data analysis tool that is utilized for cloud computation. Hail was originally built for genetics analysis, but the Table/MatrixTable/DataFrame structure and the built-in statistical analysis methods are applicable to a broader range of structured data. Our query language functions as a Python wrapper for various Hail commands. By running our query language, users send jobs to our execution system that is described in methods 1. Since the information exchange between users' local environment and our execution system is done only for the summary statistics

data, there is no risk of privacy violation. The method can be parallelized and is highly scalable and flexible (Table 1).

2.3. Realistic synthetic data generation from the query

To generate realistic synthetic (Algorithm 1) data, we first projected the real data into a low dimensional space using principal component analysis (PCA). After projecting the real data onto a four dimensional PC space, we randomly selected one sample ($S1$) and computed the pairwise distance between $S1$ and all other samples using the following distance metric:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix} \quad \mathbf{x_i} = \begin{pmatrix} PC1_i \\ PC2_i \\ PC3_i \\ PC4_i \end{pmatrix}$$

$$D_{i,j} = \left| \Lambda \cdot \left(\mathrm{x_i} - \mathrm{x_j} \right) \right|_{\ell_1}$$

Here $\mathbf{\Lambda}$ is the diagonal vector of the eigenvalues obtained from PCA, and **xi** is the projection of sample i onto each principal component (PC). The number of PCs for low-dimension projection was determined based on the variance explained by each PCs (Figure S1), and this parameters can be modified depending on the properties of the original dataset. We obtained the K nearest neighbors (we initialized K=5) to $S1$ and aggregated $S1$ and its neighbors to construct a single synthetic data sample. Aggregation was performed by randomly sampling each feature from $S1$ and the nearest neighbors. This aggregation strategy can handle missing data and ensured that the data-types remained consistent without the requirement of manual curation. This process was repeated to construct a final synthetic dataset of 773 individuals, matching the sample size of the original “real” data.

Algorithm 1 Synthetic Data Generation

Use PCA to generate *proj data* from *real_data*

for $i \in [0, num\_synthetic\_samples]$ **do**

  sample i ← CreateSyntheticIndividual (*proj_data*, *real_data*)

  store *sample_i* in Dataframe

function CreateSyntheticIndividual(*proj_data, real_data*)

  proj_sample ← Randomly sample one point from *proj data*

  proj_neighbors ← Obtain K-NN from *proj_sample*

  real_sample ← Find real samples (from *real data*) corresponding to the *proj_sample*

  real_neighbors ← Find real samples corresponding to the set of *proj_neighbors*

  synthetic_vector ← **none**

  for each feature $j \in$ *real_sample* **do**

    f eature_vals ← Feature *j* from *real_sample* and *real_neighbors*

Algorithm 1 Synthetic Data Generation

synthetic vector.append(random sample from *f eature_vals*)

return *synthetic_vector*

3. Results

3.1 Test data preparation

To demonstrate the performance of our proposed model, we used a realistic clinical trial dataset for a study testing the effect of Imatinib, an FDA approved protein-tyrosine kinase inhibitor17,18 on gastrointestinal stromal tumors (GIST).19 Specifically, the data records the results of a phase III, double blinded placebo-controlled clinical trial that was designed to test the efficacy and the optimal dosage of imatinib for GIST treatment. The dataset contains information such as drug exposure level, patient demographics (e.g. age, gender, region), baseline disease status (e.g. tumor size, location), laboratory measurements (e.g. creatinine level, white blood cell count), treatment status (e.g. start/end of treatment / placebo or drug treated), and records of recurrence or other adverse events. We labelled this dataset as *real data* and stored it in the cloud to simulate a realistic situation where public users cannot directly access the data contents, but can use our method to answer research questions. For the remainder of this manuscript, we never store the *real data* locally. Only the summary statistics obtained via our query language and the synthetic data are locally stored.

3.2. Synthetic data evaluation

Here we show that our synthetic data, while anonymized, preserves the basic properties of the real data (Figure 2A). In order to quantitatively understand the overall similarity between synthetic and real data, we performed the Kolmogorov-Smirnov test20 (KS test) for thirteen continuous features (e.g. BMI, drug exposure duration, creatinine level). KS test uses a distance metric (KS distance) determined by the the supremum-distance between two empirical distributions. The largest KS distance obtained was 0.044 indicating that the synthetic data distributions closely mimic the real data distributions. Furthermore, the lowest p-value obtained was 0.42 and therefore, we fail to reject the null hypothesis that the real and synthetic data are drawn from the same distribution. We also calculated the fraction of positively labeled samples in the real and synthetic data, for 45 binary features (e.g. gender, disease recurrence, patient death), each with more than 100 non-missing values (In other words, there are at least 100 patients who have that feature available in their clinical data record). There was a nearly perfect correlation between the fraction of positively labeled samples in the real and synthetic data (pearson's $r = 0.997$, $p < 10^{-57}$; Figure 2B). We also compared our method with two additional synthetic data generation methods.21,22 While different methods displayed variable accuracy results for different features (Figure S2), our method produced the lowest KS distance for $> 50\%$ of the features we measured (Table S1), best matching the distribution of the original, real dataset. These results suggest that our method constantly produces a high-quality proxy of the real data.

3.3. Querying language efficiency evaluation

Here we show that our querying language is scalable. Our method returns the summary statistics from a large synthetic dataset in a reasonable amount of time (less than 10 seconds for nearly 1 GiB), an order of magnitude faster than running the same analysis locally without cloud computation (more than 2 minutes for less than 1 GiB). The runtime difference becomes larger as the data size increases, reflecting the effectiveness of cloud-based parallel computing23,24 (Figure 3).

3.4. Clinical inference using synthetic and real data

Here we describe how our model can be used for biological discovery without directly viewing real clinical data, and therefore maintaining patient privacy. Specifically, we reconfirm that treatment of Imatinib is significantly associated with reduced risk of GIST recurrence17 (p =0.05). We also validate that the designed dosage level (400 mg/day) was safe for most patients on the basis of adverse events.

As a first step, we generated synthetic data and performed a two-sided Fisher exact test on this data to compare the GIST recurrence rate between the Imatinib treated (N = 273) and placebo treated (N = 331) patient populations. The Fisher exact test returned a significant p-value (p = 0.01, odds ratio (OR)= 0.54 for treated population; Table 2). To assess whether Imatinib dosage is safe for patients, we also compared the adverse event occurrence rate. We confirmed that there is no significant increase in the occurrence rate between the drug-treated population and placebo population (p = 0.33, OR = 1.32 for treated population; Table 3).

To confirm that these results obtained by analyzing the synthetic data are not due to artifacts, we queried the real data, performed the same analyses, and downloaded the resulting summary statistics. As shown in Table 2 and Table 3, the results of the statistical tests were consistent with the synthetic data. In summary, we demonstrated an example of using synthetic data to generate a hypothesis, and confirming this hypothesis by querying the real data summary statisticswithout accessing the individual level data.

4. Conclusion

While having the same concerns as sharing EHR and genomic data, sharing clinical trial data presents an additional set of challenges. A majority of patients enrolled in clinical trials have exhausted all approved treatment options. Since these individuals represent outliers in the general patient population, they are more susceptible to reidentification as well as discrimination because of their serious health conditions. Releasing such data can also spark debates about the parameters used during clinical trials, even though they have been carefully approved by ethics boards.

We have implemented a novel model for sharing clinical trial patient data while ensuring that patient privacy is protected. As our system is compartmentalized, investigators can omit specific fields in the synthetic dataset or use another synthetic data generation algorithm.

Incorporating an advanced querying language and providing corresponding synthetic data allows for both exploratory and hypothesis driven research. Synthetic data can directly be used to train machine learning and probabilistic models that aim to understand the underlying structure of the data. Once researchers have developed a new model or made a potential discovery, they can request access to the real data through a variety of clinical trial data sharing platforms such as Vivli.25,26 In addition to its flexibility, this model allows users to perform statistical tests directly on the real data via our querying language without having a locally downloaded copy. This model is also scalable and can quickly run statistical calculations on large sample sizes via parallel processing in the cloud.

We demonstrate how users can employ both real data summary statistics and synthetic data to analyze clinical trial datasets. Our synthetic data algorithm generates samples from nearly an identical distribution as the real data. Despite our success, synthetic data generation, in general, can never fully recapitulate real datasets, especially those with outlier cases. There is also variability in the synthetic data generator accuracy within and across datasets (Figure S2; Table S1). Therefore, the described system also enables clinical trial investigators to choose their own application specific synthetic data generator. For further improvements, we aim to expand the flexibility of the query system to detect potential abuse (i.e. user submitting a large volume of queries to infer concealed information) or providing users with feedback on how to improve query searches. We also envision making our model more accessible (e.g. through different cloud-computing platforms) and creating a user-friendly interactive web browser to submit queries.

We acknowledge that stakeholders, particularly owners of clinical trial data, may be reluctant to share their data, even under restricted “read only” access because of general risk aversion and potential misuse of data. We hope that current and future work on increasing the flexibility of our system, providing stakeholders with control to omit specific fields or query conditions or change access depending on the constituent (i.e. clinicians, researchers, the public, etc.), will alleviate their reluctance to share data. Followup key informant surveys is an essential future step to assess the willingness of healthcare professionals, stakeholders, and institutions to work with such a system. In summary, we hope the following dual system can spark new advances in medicine and increase the usefulness of clinical trial datasets by enabling public data-sharing while protecting patient privacy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

5. Acknowledgements

We would like to thank Vivli(https://vivli.org/) and Replica Analytics for providing us the realistic clinical trial data. We would also like to thank Lucy Mosquera for parsing the data, and Rebecca Li for providing meaningful advice. Cloud computing resources were provided by the Broad Institute of MIT and Harvard.

References

1. Taichman DB, Backus J, Baethge C, Bauchner H, De Leeuw PW, Drazen JM, Fletcher J, Frizelle FA, Groves T, Haileamlak A et al., Sharing clinical trial data: a proposal from the international

committee of medical journal editors, Annals of internal medicine 164, 505 (2016). [PubMed: 26792258]

2. Hudson KL and Collins FS, Sharing and reporting the results of clinical trials, Jama 313, 355 (2015). [PubMed: 25408371]
3. Eichler H-G, Abadie E, Breckenridge A, Leufkens H and Rasi G, Open clinical trial data for all? a view from regulators, PLoS medicine 9, p. e1001202 (2012).
4. Law MR, Kawasumi Y and Morgan SG, Despite law, fewer than one in eight completed studies of drugs and biologics are reported on time on clinicaltrials. gov, Health Affairs 30, 2338 (2011). [PubMed: 22147862]
5. Vickers AJ, Whose data set is it anyway? sharing raw data from randomized trials, Trials 7, p. 15 (2006). [PubMed: 16704733]
6. Zarin DA and Tse T, Moving toward transparency of clinical trials (2008).
7. Saito H and Gill CJ, How frequently do the results from completed us clinical trials enter the public domain?-a statistical analysis of the clinicaltrials. gov database, PLoS One 9, p. e101826 (2014).
8. El Emam K, Rodgers S and Malin B, Anonymising and sharing individual patient data, bmj 350, p. h1139 (2015).
9. OUzuner ¨, Luo Y and Szolovits P, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association 14, 550 (2007). [PubMed: 17600094]
10. Sweeney L, Replacing personally-identifying information in medical records, the scrub system., in Proceedings of the AMIA annual fall symposium, 1996.
11. Meystre SM, Friedlin FJ, South BR, Shen S and Samore MH, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC medical research methodology 10, p. 70 (2010). [PubMed: 20678228]
12. Benitez K and Malin B, Evaluating re-identification risks with respect to the hipaa privacy rule, Journal of the American Medical Informatics Association 17, 169 (2010). [PubMed: 20190059]
13. Appari A and Johnson ME, Information security and privacy in healthcare: current state of research, International journal of Internet and enterprise management 6, 279 (2010).
14. Grouin C and Zweigenbaum P, Automatic de-identification of french clinical records: comparison of rule-based and machine-learning approaches., in MedInfo, 2013.
15. Kotsiantis SB, Zaharakis I and Pintelas P, Supervised machine learning: A review of classification techniques, Emerging artificial intelligence applications in computer engineering 160, 3 (2007).
16. Choi E, Biswal S, Malin B, Duke J, Stewart WF and Sun J, Generating multi-label discrete patient records using generative adversarial networks, arXiv preprint arXiv:1703.06490 (2017).
17. Deshaies I, Cherenfant J, Gusani NJ, Jiang Y, Harvey HA, Kimchi ET, Kaifi JT and Staveley-O'Carroll KF, Gastrointestinal stromal tumor (gist) recurrence following surgery: review of the clinical utility of imatinib treatment, Therapeutics and clinical risk management 6, p. 453 (2010). [PubMed: 20957137]
18. Sarlomo-Rikala M, Kovatich AJ, Barusevicius A and Miettinen M, Cd117: a sensitive marker for gastrointestinal stromal tumors that is more specific than cd34., Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc 11, 728 (1998).
19. Nilsson B, Bümming P, Meis-Kindblom JM, Odén A, Dortok A, Gustavsson B, Sablinska K and Kindblom L-G, Gastrointestinal stromal tumors: the incidence, prevalence, clinical course, and prognostication in the preimatinib mesylate era: a population-based study in western sweden, Cancer 103, 821 (2005). [PubMed: 15648083]
20. Massey FJ Jr, The kolmogorov-smirnov test for goodness of fit, Journal of the American statistical Association 46, 68 (1951).
21. Nowok B, Raab GM, Dibben C et al., synthpop: Bespoke creation of synthetic data in r, J Stat Softw 74, 1 (2016).
22. Buciluˇa C, Caruana R and Niculescu-Mizil A, Model compression, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.
23. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ et al., Apache spark: a unified engine for big data processing, Communications of the ACM 59, 56 (2016).

24. Jadeja Y and Modi K, Cloud computing-concepts, architecture and challenges, in 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012.

25. Li R, Scott J, Rockhold F, Sim I, Wood J et al., Moving data sharing forward: The launch of the vivli platform (2018).

26. Bierer BE, Li R, Barnes M and Sim I, A global, neutral platform for sharing trial data, New England Journal of Medicine 374, 2411 (2016). [PubMed: 27168194]

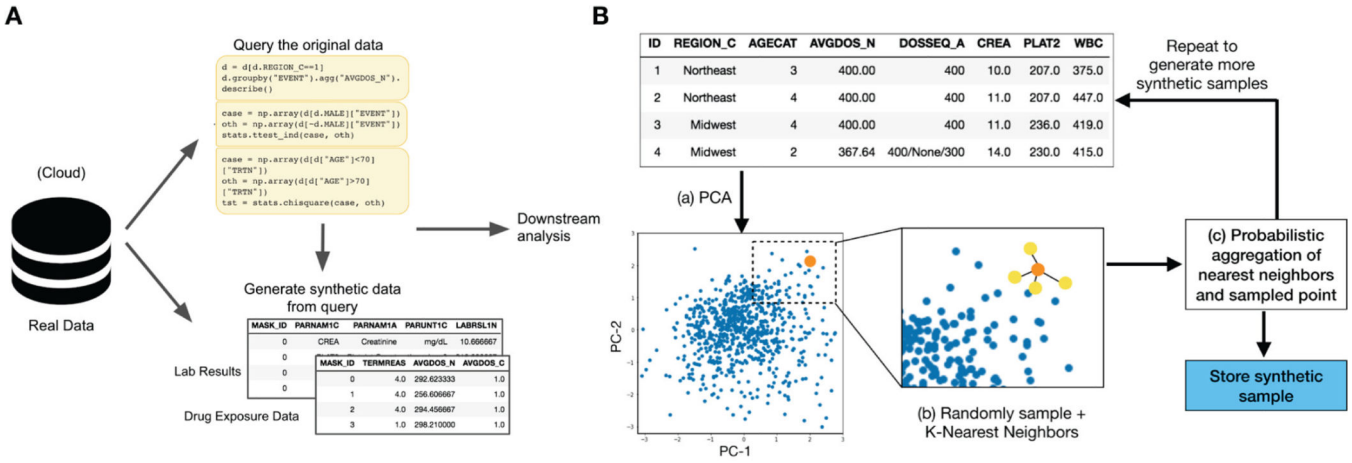

Figure 1.

Graphical overview of the data sharing model. **(A)** Real data is stored securely on the cloud and can be accessed by user queries. Synthetic data matching the specified query is also generated and returned to the user for downstream analysis. **(B)** Synthetic data for the model presented in panel A was obtained by first performing PCA to reduce the dimensionality of real data(a). One sample from the projected data was randomly obtained and the nearest k-neighbors for that sample was determined using a PC weighted distance metric(b). Aggregating data from the sample and its nearest neighbors was used to construct a synthetic sample(c). This process was repeated to generate 773 synthetic patients.

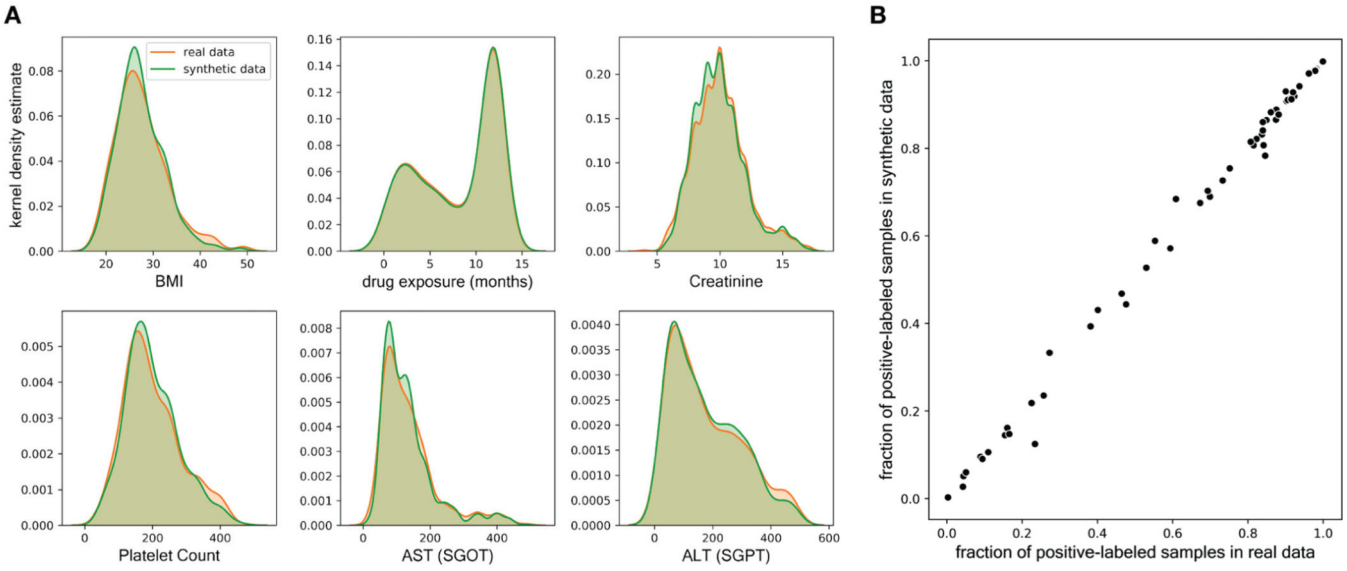

Figure 2.

Comparing the real clinical trial dataset and the corresponding synthetically generated dataset. (**A**) Real data distribution (orange) and synthetic data distribution (green) of six clinical features. Distributions were obtained from the empiricle data via kernel density estimation **(B)** Comparison of the fraction of positively labeled data between the real and synthetic datasets, for 45 binary features with more than 100 non-missing values.

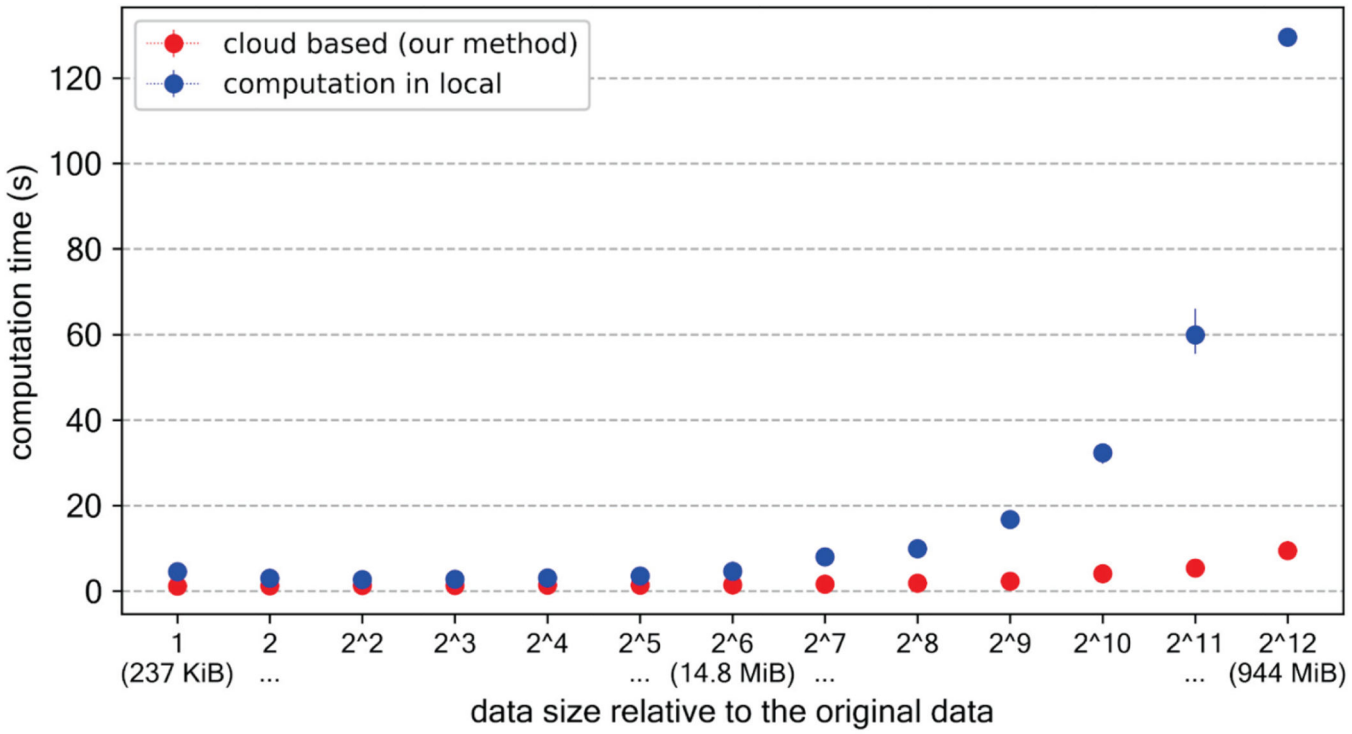

Figure 3.

The runtime comparison between our cloud based model versus running computation locally. The runtime is measured as the number seconds required for aggregating the data and creating a contingency table

Table 1.

Querying language examples

operation	example
view available features	df.describe()
filtering	df.filter(feature A >10)
grouping	df.groupby(feature B)
aggregation	df.groupby(feature B).aggregate(feature C).summarize(max, min, mean, median, sd)
correlation	df.pearson corr(feature A, feature B) statistical tests df.t test(feature A, feature B) visualization df.hist(feature A)
combination of above	df.filter(feature A>10).groupby(feature B).t_test(feature C, feature D)

Table 2.

Contingency table for real and synthetic data: Comparing the recurrence rate for drug treated and placebo population

	synthetic data			real data	
	recurrence	**no recurrence**		**recurrence**	**no recurrence**
drug treated	27	246		30	267
placebo	56	275		47	255
(p-value, OR)	(0.01, 0.54)			(0.05, 0.61)	

Table 3.

Contingency table for real and synthetic data: Comparing the adverse event rate for drug treated and placebo population

	synthetic data			real data	
	adverse event	**no adverse event**		**adverse event**	**no adverse event**
drug treated	30	215		34	252
placebo	29	274		40	258
(p-value, OR)	(0.33, 1.32)			(0.62, 0.87)	