

UC San Diego

UC San Diego Previously Published Works

Title

Collaborative privacy-preserving analysis of oncological data using multiparty homomorphic encryption.

Permalink

<https://escholarship.org/uc/item/2008k178>

Journal

Proceedings of the National Academy of Sciences of USA, 120(33)

Authors

Geva, Ravit
Gusev, Alexander
Polyakov, Yuriy
[et al.](#)

Publication Date

2023-08-15



DOI

10.1073/pnas.2304415120

Peer reviewed



Collaborative privacy-preserving analysis of oncological data using multiparty homomorphic encryption

Ravit Geva^{a,1} , Alexander Gusev^{b,1}, Yuriy Polyakov^{c,1} , Lior Liram^{c,1}, Oded Rosolio^{c,1}, Andreea Alexandru^{c,1}, Nicholas Genise^{c,1}, Marcelo Blatt^{c,1}, Zohar Duchin^{c,1}, Barliz Weissengrin^{a,1}, Dan Mirelman^{a,1}, Felix Bukstein^{a,1}, Deborah T. Blumenthal^{a,1}, Ido Wolf^{a,1}, Sharon Pelles-Avraham^{a,1}, Tali Schaffer^{a,1}, Lee A. Lavi^{a,1}, Daniele Micciancio^{c,d,1}, Vinod Vaikuntanathan^{c,e,1}, Ahmad Al Badawi^{c,1}, and Shafi Goldwasser^{c,f,1,2}

Contributed by Shafi Goldwasser; received March 17, 2023; accepted June 9, 2023; reviewed by Abhishek Jain and Benny Pinkas

Real-world healthcare data sharing is instrumental in constructing broader-based and larger clinical datasets that may improve clinical decision-making research and outcomes. Stakeholders are frequently reluctant to share their data without guaranteed patient privacy, proper protection of their datasets, and control over the usage of their data. Fully homomorphic encryption (FHE) is a cryptographic capability that can address these issues by enabling computation on encrypted data without intermediate decryptions, so the analytics results are obtained without revealing the raw data. This work presents a toolset for collaborative privacy-preserving analysis of oncological data using multiparty FHE. Our toolset supports survival analysis, logistic regression training, and several common descriptive statistics. We demonstrate using oncological datasets that the toolset achieves high accuracy and practical performance, which scales well to larger datasets. As part of this work, we propose a cryptographic protocol for interactive bootstrapping in multiparty FHE, which is of independent interest. The toolset we develop is general-purpose and can be applied to other collaborative medical and healthcare application domains.

multiparty fully homomorphic encryption | privacy-enhancing technologies | oncology | privacy-preserving data collaboration

There is a growing recognition of the important contribution of real-world data (RWD) in supporting healthcare decision-making in general (1, 2) and specifically in oncology (3, 4). RWD are routinely collected from a variety of sources, such as electronic health records; medical claims and billing data; product and disease registries; and mobile devices (5). RWD can complement data generated from randomized control trials (RCTs). While RCTs analyze data collected from controlled, limited, and homogenous patient populations, RWD allow the evaluation of larger and broader-based patient populations within the context of routine clinical practice (6). Sharing RWD between several data owners results in a more complete dataset than that obtained from a single data source and thus allows broader data analyses for better decision-making (7, 8). In addition, healthcare data can be viewed as a revenue-producing asset that can be monetized. RWD analysis can save costs to the pharmaceutical industry by improving the identification of target populations, endpoints, and inclusion criteria, and thereby the overall study design (9). Some of the main challenges of using RWD for healthcare decision-making are the facts that healthcare data are fragmented and originate from multiple sources and that stakeholders are frequently hesitant to share or integrate their data, mainly due to trust issues (10).

Generally, patient data may be shared only if the patient's consent had been obtained for a given purpose or if the data are anonymized or deidentified (11). While patients participating in RCTs can give their consent for data sharing, patients for whom RWD are collected do not necessarily provide their consent in advance for this purpose. In such cases, data anonymization is required; however, anonymization procedures are recognized as being time-consuming, requiring manual intervention that can result in human error, difficult to scale, and challenging in terms of assessing their results (12). Furthermore, anonymization requires the removal of sufficient patient data to prevent any possible re-identification, many times resulting in the impairment of scientific analysis and utility (13). Healthcare data challenges, particularly patient privacy, data ownership, and data fragmentation, call for a data collaboration technology that, on the one hand, allows different parties to share their data and analytics, and, on the other hand, protects patient privacy and data ownership.

The two common cryptographic approaches to share and analyze sensitive data without compromising patient privacy and data ownership are secure MultiParty

Significance

Improving clinical decision-making and research-based patient treatment relies on access to comprehensive clinical datasets obtained by sharing real-world healthcare data. However, without guaranteed patient privacy, proper protection of datasets, and control over data usage, stakeholders withhold their data from inclusion in larger clinical datasets. Fully homomorphic encryption (FHE) is a cryptographic tool that can address these issues by enabling computation on encrypted data without ever decrypting the raw data or intermediate results. We develop a general-purpose toolset for collaborative privacy-preserving analytics, including survival analysis, logistic regression training, and several common descriptive statistics, using multiparty FHE. We exemplify our toolset performance over encrypted oncological data and emphasize that it applies to other collaborative medical and healthcare application domains.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

¹R.G., A.G., Y.P., L.L., O.R., A.A., N.G., M.B., Z.D., B.W., D.M., F.B., D.T.B., I.W., S.P.-A., T.S., L.A.L., D.M., V.V., A.A.B., and S.G. contributed equally to this work.

²To whom correspondence may be addressed. Email: shafi@csail.mit.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2304415120/-/DCSupplemental>.

Published August 7, 2023.

Computation (MPC) and Fully Homomorphic Encryption (FHE).^{*} Both allow performing computations over encrypted data, but the underlying mechanisms are different. MPC, which was introduced by Yao (14), uses an approach where each party holds a secret, and they perform computations on masked data using an interactive protocol. MPC is communication-bound and typically based on either garbled circuits or secret-sharing schemes (15). FHE, which was first achieved by Gentry (16), provides a noninteractive mechanism for performing computations on encrypted data in an untrusted environment, without ever decrypting the data or intermediate results. Only once the final computation results are obtained, the decryption of results may be performed by a different party that has the underlying secret key. FHE is compute-bound and typically based on lattice cryptography, which is resistant to attacks by quantum computers (17, 18).

Notable recent studies on privacy-preserving analysis of individual-level healthcare data using MPC include Cho et al. (19) and Hie et al. (20). Cho et al. (19) report on large-scale genome-wide analysis of genotypic and phenotypic data using MPC. They perform a genome-wide associate study (GWAS) by dividing data among multiple servers and computing the GWAS via MPC among the servers. They demonstrate that their results provide adequate accuracy, and reasonable runtime (about 37 h) can be achieved for problem sizes of 100,000 individuals and 500,000 single nucleotide polymorphisms, enabling real-scale privacy-preserving GWAS. Hie et al. (20) develop a computational protocol for securely training a predictive model of drug–target interactions on a pooled dataset using MPC. Their protocol for neural network training runs within days on a real dataset of more than one million interactions and is more accurate than state-of-the-art drug–target interaction prediction methods.

FHE has also seen significant success in performing privacy-preserving analysis for certain healthcare use cases. Note that almost all FHE results described below are based on the Cheon–Kim–Kim–Song (CKKS) FHE scheme (21), which is the most efficient scheme for real-number arithmetic and many machine learning applications (18). For instance, Blatt et al. (22) demonstrate that FHE can perform GWAS for 100,000 individuals and 500,000 single nucleotide polymorphisms in less than 6 h, hence achieving a better runtime than the prior MPC approach of Cho et al. (19) while still providing a comparable accuracy. Kim et al. (23) were able to train a logistic regression model using an encrypted dataset for 1,579 individuals with 18 binary genotypes and a binary phenotype outcome (cancer/no cancer). Using several aggressive approximations and optimized values of tunable parameters, the authors were able to perform encrypted logistic regression training in about 6 min on a commodity desktop machine.

However, practical results with FHE can typically be achieved only for relatively shallow (limited-depth) computations that do not require bootstrapping, a special procedure that refreshes exhausted ciphertexts to enable more computations. Bootstrapping is a computationally expensive and memory-intensive procedure that needs to be invoked many times for deep computations such as logistic regression training or deep neural network inference. In applications with bootstrapping, the FHE runtimes and memory consumption become much higher. Notable recent studies implementing machine learning capabilities using CKKS bootstrapping are Han et al. (24) and

Lee et al. (25). Han et al. (24) present a logistic regression training capability based on FHE that can train a model with 422,108 samples over 200 features in about 17 h. Lee et al. (25) develop a privacy-preserving CNN inference solution that can classify with a ResNet-56 model a CIFAR-10 image in about 2 h. In both cases, most of the computation time is spent on CKKS bootstrapping.

To minimize the number of CKKS bootstrapping invocations in applications of FHE, researchers often use hand-tuned low-accuracy–low-degree approximations for nonlinear functions and dataset-optimized parameters, e.g., learning rate, which allows to significantly improve the efficiency of an FHE computation for a given dataset. But as soon as the FHE solution is applied to other datasets, the solution stops working correctly or achieving adequate accuracy. For example, Han et al. (24) used a degree-3 polynomial approximation of the sigmoid function obtained using the least squares method for the range of $[-8, 8]$. Our analysis of polynomial sigmoid approximations in the Nesterov gradient descent method of logistic regression training (same method as in ref. 24) for another large dataset shows that a Chebyshev interpolation in the range of $[-32, 32]$ using a polynomial of degree of at least 32 is needed to achieve satisfactory accuracy results (see *SI Appendix* for details). Generally speaking, both the range and polynomial degree may significantly vary from one dataset to another. If a more costly polynomial approximation is used, the bootstrapping has to be invoked much more frequently. For comparison, the logistic regression solution in ref. 24 performed bootstrapping every 5 iterations whereas ours calls bootstrapping after each iteration.

To address the FHE bootstrapping inefficiency, Froelicher et al. (26) present an interactive computation framework based on multiparty FHE (the algorithms were originally introduced in ref. 27), which uses FHE for most of the computations and interactive techniques for bootstrapping and several other operations. In multiparty FHE [typically referred to as threshold FHE in cryptography literature (28)], each party may have a secret share (similar to classical MPC based on secret sharing), and distributed key generation and decryption protocols are executed involving all parties with secret shares (Fig. 1). The main efficiency benefit of this approach as compared to FHE is that bootstrapping can be done interactively much faster (by two orders of magnitude or even more) than in the classical FHE setting. The authors demonstrate the use of their privacy-preserving framework for Kaplan–Meier survival analysis in oncology and genome-wide association studies in medical genetics. Froelicher et al. (26) consider the federated collaboration model between data owners, where each party contributes a subset of records to the full dataset used for privacy-preserving analysis (Fig. 2A).

Our work extends and improves the multiparty FHE framework of ref. 26 in several different ways.

First, we add the private join collaboration model where multiple parties can contribute data for the same records (e.g., individuals) in a way where the data owners do not learn which records match (with only the computation party learning the intersection size in the case of two data owners), and these joined data are then used for further analysis using multiparty FHE (Fig. 2B).

Second, we introduce a more efficient interactive bootstrapping procedure for the case of two parties and improve the more general (for any number of parties) interactive bootstrapping method used in ref. 26, and initially proposed in ref. 27.

Third, we extend the list of computations to provide a more general toolset for the privacy-preserving analysis of oncological

^{*}Our paper uses a number of specialized terms in cryptography and oncology; for convenience, we provide a glossary of these terms in *SI Appendix, Table S21*.

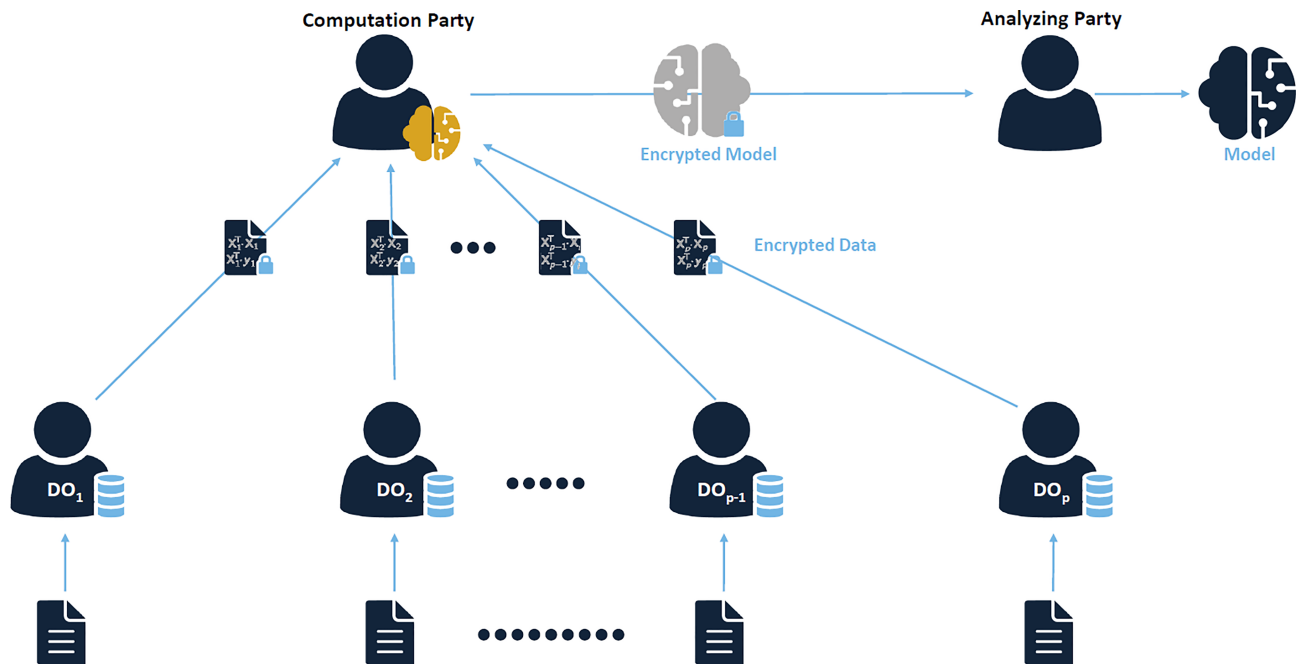


Fig. 1. Schematic of multiparty (threshold) FHE. Any party may have a secret share (assignment of secret shares is determined by the use case). At least two parties have secret shares. First, all parties with secret shares perform distributed key generation to compute the common public key, corresponding to the sum of secret shares. Next, the data are encrypted by each data owner (DO) using the common public key. Then, the computation is performed by the computation party (CP). If interactive bootstrapping is needed, the CP interacts with the parties that have secret shares. Finally, the encrypted result is decrypted using a distributed decryption procedure involving all parties with secret shares. The analyzing party (AP) is the party that gets to see the result of the computation and can be the same as one of DOs (multiple DOs may serve as APs in some use cases). In the setting of multiparty FHE, the CP can be one of the DOs. The DOs, CP, and AP are separated in the schematic to show all possible roles involved in the multiparty FHE collaboration model.

data. The computations implemented in our work include mean, median, SD, frequency, χ^2 test, t test, survival analysis (Kaplan–Meier plots and log-rank test), and logistic regression training over encrypted data.

Results

We applied our multiparty FHE toolset to two different oncological datasets: a real-world dataset of colorectal cancer patients' survival data at the Tel Aviv Sourasky Medical Center and a previously published dataset based on two clinical trials of immunotherapy in renal cell carcinoma (29).

The real-world dataset of colorectal cancer patients' survival data includes 623 patients and 24 variables, amounting to 14,952 items of data. The goal of the study was to examine the effect of oxaliplatin treatment with and without cannabis for patients with colorectal cancer. Statistical analysis of key oncological endpoints was blindly performed on both the raw data and FHE-encrypted data using descriptive statistics and survival analysis with Kaplan–Meier curves and log-rank tests. The results were then compared with an accuracy goal of two decimals. Early results of this study (for the single-key FHE setting) are reported in ref. 30. The study included the following statistical analyses: mean, median, and SD for the age of cancer onset; frequency analysis for sex; χ^2 -test between cannabis indicator (with or without cannabis) and diagnosis, χ^2 -test between cannabis indicator and sex; t -test for cannabis indicator by age of onset. Kaplan–Meier and log-rank survival analysis was performed to examine the effect of the treatment with cannabis on the overall survival of patients.

All accuracy metrics were found to be within the predetermined accuracy goal of two decimal digits. The Kaplan–Meier curves for both the data in the clear and encrypted data are illustrated in Fig. 3. The numerical results of the first 15 wk out

of 141 wk following the first oxaliplatin treatment are listed in *SI Appendix, Table S1*. The runtime of less than half a minute was observed for descriptive statistics and about 3 min for the survival analysis (30). Note that the time of the anonymization and statistical analyses performed on the raw dataset by a statistician, the method commonly used in clinical oncology, is estimated to be about 10 h, which is significantly higher than the runtime of FHE computations. As this dataset is not publicly available (see the *Data, Materials, and Software Availability* for more details), we performed a similar analysis for a publicly available dataset so that our results could be independently reproduced. We further extended the analysis to include logistic regression training, another useful tool for oncological and broader healthcare studies.

Next, we show an example of applying our multiparty FHE toolset to an analysis of a previously published dataset, providing detailed results for it in *SI Appendix*. Individual-level data from two clinical trials of immunotherapy in renal cell carcinoma were accessed from prior publications (29). In brief, a PD-1 immune checkpoint inhibitor (nivolumab) was evaluated for 1,006 patients with advanced clear-cell renal cell carcinoma (ccRCC) in the CheckMate 025 and CheckMate 010 randomized clinical trials, as compared to the standard of care with mTOR inhibitors (everolimus). Clinical outcome data included overall survival and progression-free survival as well as basic patient demographics, and genomic data included tumor whole exome sequencing. Prior work had identified a survival benefit for nivolumab as well as improved progression-free survival for the subset of patients with mutations in PBRM1, and we focused on these positive controls in our analyses here.

First, we evaluated the accuracy of basic demographic summaries of age, sex, prior treatment, and objective response rate (ORR) within and across the treatments. More concretely, we computed the mean, median, and SD for age; performed

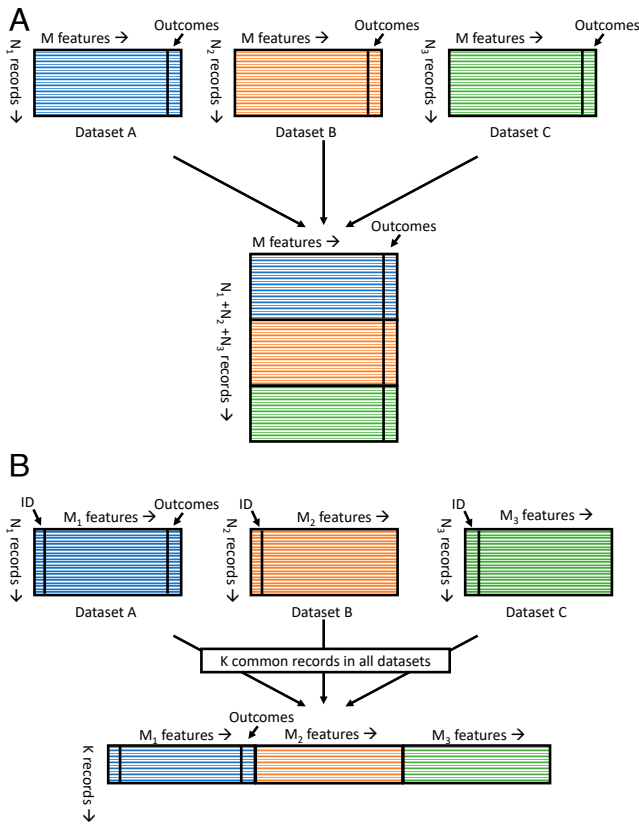
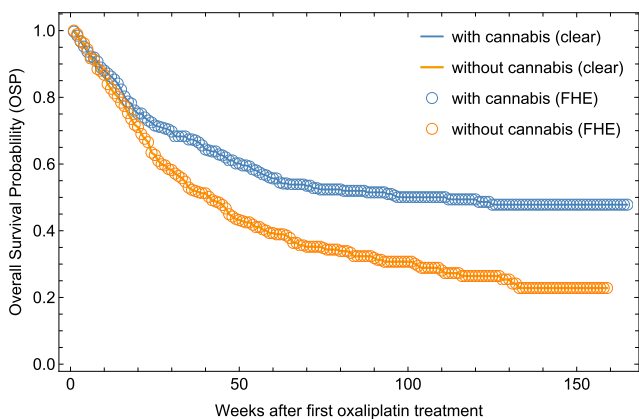


Fig. 2. Collaboration models for privacy-preserving analysis of data from multiple DOs. (A) Federated model: each party contributes a subset of records to the full data set used for privacy-preserving analysis. This model supports two scenarios: 1) local FHE computations are performed by each DO (similar to the federated learning setting) and 2) an FHE computation is carried out by CP on the stacked encrypted data set. (B) Private join model: multiple parties contribute features data for the same records in a way where the parties do not learn which records match. Then an FHE computation is performed on the linked encrypted data.

a χ^2 -test between ORR and trial arm, where trial arm was set to 1 for nivolumab and 0 for everolimus; performed *t*-tests for age by trial arm (*t*-test 1) and age by ORR groups (*t*-test



A Kaplan-Meier curves: in the clear vs FHE. The number of patients at risk:

Time (weeks)	25	75	125	175
With cannabis	239	138	78	58
Without cannabis	180	90	32	18

2); we evaluated the frequency for sex (frequency 1), benefit (frequency 2), PBRM1 (frequency 3), and the number of prior therapies (frequency 4).

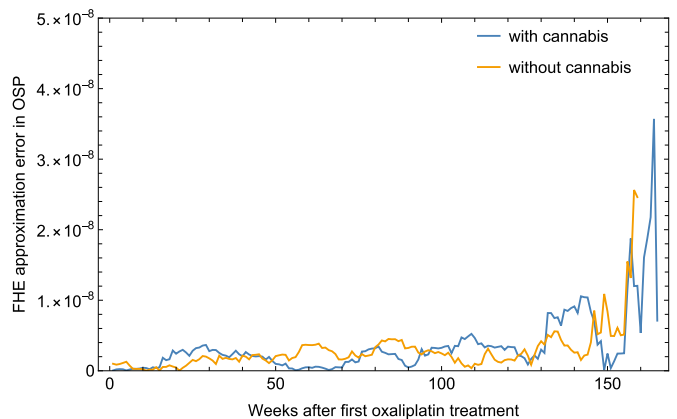
Second, we conducted survival analyses where mortality was the endpoint and patients were censored at loss-to-follow-up, with statistical significance assessed by log-rank test and Kaplan–Meier analysis. For the treatment arm positive control (nivolumab vs. other), which corresponds to Kaplan–Meier and log-rank scenario 1, we observed a significant association, e.g., the *P*-value for the log-rank test was 0.001. For the sex-stratified negative control, which corresponds to Kaplan–Meier and log-rank scenario 2, as expected, we observed no significant difference between groups, e.g., the *P*-value for the log-rank test was 0.104.

Third, we conducted biomarker survival analyses where progression-free survival was the endpoint and patients were censored at loss-to-follow-up. For the positive control within the nivolumab arm, which corresponds to Kaplan–Meier and log-rank scenario 3, patients with PBRM1 mutations exhibited significantly longer survival than noncarriers by long-rank test, e.g., the *P*-value for the log-rank test was 0.006.

Fourth, we conducted a logistic regression analysis where ORR was the outcome, and age, sex, and trial arm were independent variables. For this analysis, ORR was defined as 1 for complete response or partial response (CR/PR) and 0 for stable disease or progression disease (SD/PD). As expected, a significant association was observed with the trial arm.

For the multiparty FHE experiments, the full dataset was filtered down and broken into different subsets to emulate realistic private join scenarios with two data owners (see *SI Appendix, Table S2* for details). Note that the runtime and communication costs for the private join protocol are negligibly small as the numbers of records and features for the oncological dataset are not high. Hence, the runtimes reported here are determined by the FHE computations performed after executing the join protocol.

Table 1 shows the relative errors for descriptive statistics and survival analysis, as compared to the results in the clear. For all computations, accuracy of more than 5 decimal digits (as compared to the computations in the clear) was achieved. Note that for frequency computations, the error was zero because we



B FHE approximation error computed as the absolute value of the difference in Overall Survival Probability (OSP) between the results in the clear and FHE results. The error in OSP gradually increases with time as expected, since at each time step the number of (FHE) computations for that value increases as well. However, the error is less than 5×10^{-8} , which is negligible for the survival analysis.

Fig. 3. Kaplan–Meier survival analysis of the real-world dataset for colorectal cancer patients: results in the clear vs encrypted data (A), FHE approximation error (B). There were two groups of patients treated with oxaliplatin: The first group was taking cannabis and the other was not. The survival analysis results for the encrypted dataset were found to be accurate up to 7 decimal digits compared to the results for the unencrypted dataset (see also *SI Appendix, Table S1* for numerical details). Note that this Kaplan–Meier analysis has no clinical significance and should not be interpreted as such. The analysis was performed solely for the purpose of testing the proposed FHE method.

Table 1. Numerical accuracy for descriptive statistics and survival analysis computed with FHE vs. the computations in the clear using the data published in ref. 29

FHE Scheme	Computation	Statistic	Rel Error
CKKS	Mean	Average	2.83e-12
	SD	SD	1.62e-07
	Median	Quantile	0.0e+00
	t-test 1	t-score	1.57e-09
	t-test 2	t-score	1.60e-09
	χ^2	χ^2	1.91e-09
	Kaplan–Meier 1	Probability	2.04e-07
	Kaplan–Meier 2	Probability	7.39e-06
	Kaplan–Meier 3	Probability	2.08e-07
	log-rank 1	χ^2	3.21e-08
	log-rank 2	χ^2	4.92e-08
	log-rank 3	χ^2	3.80e-08
	BFV	Frequency 1	Count
Frequency 2		Count	0.0e+00
Frequency 3		Count	0.0e+00
Frequency 4		Count	0.0e+00

The mean, SD, and median were computed for age; t-test 1 for age by trial arm; t-test 2 for age by ORR groups; χ^2 -test for ORR by trial arm; Kaplan–Meier and log-rank 1 for overall survivability by arm (OS, OS_CNRS); Kaplan–Meier and log-rank 2 for overall survivability by sex (OS, OS_CNRS); Kaplan–Meier and log-rank 3 for progression-free survival with somatic mutations by arm (PFS, PFS_CNRS); frequency 1 for sex; frequency 2 for benefit; frequency 3 for PBRM1; frequency 4 for the number of prior therapies. All these computations did not require bootstrapping.

used BFV, an exact homomorphic encryption scheme, for these computations. Table 2 illustrates the runtime and storage performance. Besides more complex survival analysis, all computations take less than half a minute. The survival analysis takes up to 1 min and a half. The memory requirements do not exceed a few gigabytes. These results imply that privacy-preserving descriptive statistics and survival analysis using multiparty FHE are already practical for typical oncological datasets.

Our results for logistic regression training using multiparty FHE suggest that accuracy of at least 6 decimal digits was achieved after 100 iterations (*SI Appendix*). The performance results for

Table 2. Runtime and storage measurements for computations in Table 1; Kaplan–Meier and log-rank survival analysis methods are abbreviated as KM and LR, respectively; KeyGen and Comp correspond to key generation and computation

	Peak RAM [GB]	KeyGen [s]	Comp [s]	KeyGen [MB]	Comp [MB]
Mean	2.13	1.2	1.6	140	1.0
SD	2.51	6.2	2.5	600	3.0
Median	2.17	2.0	21.7	244	108.0
t-test 1	1.97	2.9	3.2	477	1.3
t-test 2	1.99	2.9	3.1	477	1.3
χ^2	2.43	9.7	6.7	1,458	1,458.0
KM & LR 1	2.36	3.1	79.3	546	3,030.0
KM & LR 2	2.39	2.7	79.6	546	3,030.0
KM & LR 3	2.43	2.9	20.7	546	727.0
Frequency 1	2.03	0.3	1.2	23	0.5
Frequency 2	2.06	0.3	1.2	23	0.8
Frequency 3	2.03	0.3	1.2	23	0.8
Frequency 4	2.13	0.3	1.2	23	1.3

Table 3. Runtime and memory performance for logistic regression training; interactive bootstrapping is performed after every iteration; KeyGen refers to key generation

Iterations	Peak RAM [GB]	Total [s]	KeyGen [s]	Iteration time [s]
10	7.752	83.4	30.6	5.3
20	7.943	128.3	30.7	4.9
30	8.072	175.0	30.5	4.8
40	8.199	222.5	30.8	4.8
50	8.257	264.9	30.7	4.7
60	8.273	314.7	31.2	4.7
70	8.630	362.4	30.6	4.7
80	8.625	414.2	31.4	4.8
90	8.703	450.0	30.7	4.7
100	8.760	500.0	30.7	4.7

logistic regression training are illustrated in Table 3. One iteration takes about 5 s, and the overall runtime of 500 s is observed for 100 iterations. The memory requirements do not exceed 9 GB. Both the runtime and memory are significantly smaller than for the scenario of noninteractive FHE bootstrapping, which makes it possible to run training on a commodity desktop or server machine. For instance, one noninteractive CKKS bootstrapping operation for this setting takes at least 24 s (31).

To evaluate the scalability of our multiparty FHE framework, we ran the most computationally intensive capability considered in our work, logistic regression training in the private join collaboration setting, for much larger problem sizes than the oncological dataset. We considered the case of two DOs, where the first DO has the encrypted outcome data and the second DO has the features data. In this case, only encrypted outcome data need to be sent to the party that sees the result of the private join, which implies that the performance cost of private join is almost negligible as compared to logistic regression training. We ran performance experiments on a server commodity system with 72 threads for simulated datasets from 16,384 samples to 1,048,576 samples (doubling each time), all with 256 features. For the sample sizes from 16,384 to 262,144, the runtime of one iteration stayed roughly the same at about 60 s. For 524,288 and 1,048,576 samples, the runtime was 98 and 187 s, respectively. This implies that our logistic regression solution scales relatively well with the number of cores of the server system; more concretely, the runtime degrades only by a factor of 3 when increasing the number of samples by a factor of 64.

Discussion

Our approach enables multiple analyses of clinical and genetic data.

First, the federated model allows multiple institutions with disjoint clinical and genomic data to perform secure joint analyses across all patients without decrypting the underlying individual-level values. Clinical trials conducted across multiple institutions can now evaluate drug efficacy by computing secure Kaplan–Meier and log-rank analyses. In particular, genomic data, which are often considered sensitive patient data that may be impossible to deidentify, can be leveraged to perform biomarker analyses within or across treatment subgroups to identify treatment modifiers. We demonstrate the feasibility and accuracy of this approach by recapitulating the protective effect of somatic PBRM1 mutations in patients on immune checkpoint

inhibitors (32). Such genetic biomarkers can prioritize effective treatments for patients who would not have otherwise received them and further improve patient outcomes as well as expand our understanding of disease etiology. In addition to analyses of survival-related outcomes, recent studies have identified biomarkers associated with drug safety outcomes through time-to-event analyses of adverse events (33). While overall adverse event rates are reported with clinical trials, individual-level data are typically not made available. Our methodology would enable secure, large-scale adverse event studies across multiple trials with the potential of identifying actionable predictors of toxicities that can be mitigated before they occur.

Second, the private join model allows investigators to conduct secure analyses of clinical and genomic data when the underlying data reside at different sites and cannot be integrated in the clear. This is often the case for biobank cohorts, which routinely collect rich genetic/genomic data but typically only sparse clinical measurements through billing codes, whereas detailed chart review and clinical records abstraction are typically conducted under strict Institutional Review Board protocols at medical institutions which do not conduct routine biobanking. For example, while the UK Biobank has collected a wide array of omics data including whole-exome and whole-genome sequencing, it only has rudimentary treatment billing codes and does not report the duration, dosage, or treatment response (34). Individual institutions that collect such data but do not conduct genotyping could thus use our approach to tap into existing genomic resources while retaining patient confidentiality across both cohorts. While genetic data were used for demonstration in this study, our approaches would naturally apply to broader classes of omics that can be summarized as continuous or categorical values, such as magnetic resonance images (35) or structural brain images (36). Our approach thus enables privacy-preserving identification of clinical biomarkers across institutions and data silos.

Our approach has several limitations and areas for future work. First, our private join approach requires sending the full encrypted DO datasets (the fields to be computed on) to the party computing the join of the datasets from DOs. This requirement can only be removed if the DOs are allowed to learn something about the intersection. Second, the extension of private join to more than two DOs prevents DOs from learning anything about the intersection only if the computing party performs expensive homomorphic rearrangement of encrypted data (using many rotations) or if the DOs are not allowed to collude with each other. Devising a more efficient solution for extending the private join to more than two DOs is left for future work. Third, while the FHE scheme we use is plausibly postquantum secure, the private join protocol in our framework uses a commutative deterministic cipher based on an elliptic-curve instantiation of the decisional Diffie–Hellman (DDH) problem, which is secure against classical computer attacks but is not quantum-resistant. As future work, we will look into developing a quantum-resistant version of this private join protocol based on lattices. See *SI Appendix* for more details on the first three limitations. Fourth, our multiparty FHE framework is based on the same adversarial model as single-key FHE, i.e., it is secure against semihonest adversaries.

Materials and Methods

Software Implementation. We implemented our multiparty FHE framework in PALISADE v1.11.9 (37), an open-source lattice cryptography software library that includes all common FHE schemes. For the experiments, we used full Residue Number System (RNS) variants of the Cheon–Kim–Kim–Song (CKKS) and

Brakerski/Fan–Vercauteren (BFV) FHE schemes (21, 38, 39), which are already available in PALISADE. The full RNS variants of CKKS and BFV are described in refs. 40 and 41, respectively.

Multiparty FHE. Our implementation is based on the threshold FHE construction proposed in ref. 28. We consider the scenario of additive secret sharing where the sum of all secret shares corresponds to the underlying secret key, but this secret key is never revealed. PALISADE provides threshold FHE extensions (without interactive bootstrapping) for the CKKS and BFV schemes (37).

Private Join. The private join collaboration model is a generalization of the private intersection-sum-with-cardinality protocol proposed in ref. 42. We extend the original protocol from encrypted summation to arbitrary encrypted computations and add support for two or more DOs with FHE-encrypted data. The model includes a CP and multiple DOs. Each DO has a subset of features for common records (all DO datasets are encrypted using FHE). The purpose of join is to link the datasets from DOs into a single dataset and perform FHE computations on it. The same record identification scheme is used for all DOs, i.e., each common record is uniquely described by the same identifier. The join is performed based on exact matches. The CP computes the intersection by interacting with the DOs, and the DOs also interact with each other. A deterministic commutative cipher based on an elliptic curve instantiation of the DDH problem is used to compute a common hash (encryption) for each record. As part of the protocol, the records get randomly shuffled and random identifiers get inserted. In the case of two DOs, the CP learns the intersection size whereas the DOs learn nothing about the intersection. The protocol and security proofs for it are described in detail in *SI Appendix*.

Interactive Bootstrapping. Our multiparty FHE framework includes two interactive bootstrapping procedures. The first algorithm achieves more efficient bootstrapping for two parties and is a contribution of our work. In contrast to the more general procedure of ref. 27 that requires at least three extra RNS residues, our algorithm requires only one extra RNS residue, which reduces the computational complexity of the full FHE workflow (not only the interactive bootstrapping invocations). Our procedure is based on a technique of distributed rounding. We describe the CKKS instantiation of our 2-party interactive bootstrapping protocol along with the security proofs in *SI Appendix*. The second algorithm supports any number of parties and is an optimized version of the procedure proposed in ref. 27. Our variant reduces the computational complexity associated with sampling: One polynomial Gaussian sampling is eliminated because it is not needed for security. We implement both protocols in PALISADE.

Experimental Test Bed. FHE computations for the oncological dataset were performed using a server with Intel(R) Xeon(R) Platinum 8275CL @ 3.00 GHz and 96 GB of RAM. The experiments were run at 16 threads using the PALISADE multithreading capability based on OMP. The server had Ubuntu 20.04 and g++ (GCC) 9.3.0 installed. All parties were connected via a local area network connection. Note that interactive computations were needed only for the logistic regression training case, where the communication requirements for interactive bootstrapping are small (a single ciphertext is transferred between the interacting parties). The scalability experiments were run on a Ubuntu 20.04 server with Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40 GHz (72 threads) and 128 GB RAM.

Computations. Before running the FHE computations, the source data were preprocessed to normalize the input data and exclude NA records. Then, analysis in the clear was performed to generate reference results for evaluating the accuracy of FHE analysis. The FHE computations for mean, SD, median, χ^2 -test, *t*-test, survival analysis (Kaplan–Meier plot and log-rank test), and logistic regression training were performed using the CKKS scheme. The CKKS in PALISADE was configured to use hybrid key switching, the scaling factor size was set to 50 bits (55 for SD), the standard mode without automatic rescaling was used, the ring dimension was automatically set to the minimum value required for achieving 128 bits of security, and the multiplicative depth was set to the minimum needed to achieve the correctness. All other parameters

were set to PALISADE defaults for CKKS. For frequency computations, we used the BFV scheme configured to use the Brakerski-Vaikuntanathan key switching and plaintext modulus of 1032193. The ring dimension was automatically set to the minimum value required for achieving 128 bits of security and the multiplicative depth was set to the minimum needed to achieve correctness. All other parameters were set to PALISADE defaults for BFV. For logistic regression training, we used the 2-party interactive bootstrapping proposed in this work. For all computations, we used a 2-DO private join setup, i.e., the full dataset was filtered down and broken into different subsets to emulate realistic private join scenarios with two DOs. More details about the computations and private join setup are provided in *SI Appendix*.

Our main experiments used a previously published dataset based on two clinical trials of immunotherapy in renal cell carcinoma (29).

Data, Materials, and Software Availability. The PALISADE version we used is publicly available in GitLab at <https://gitlab.com/palisade/palisade-release> (37). The code with the FHE computations, including descriptive statistics, survival analysis, and logistical regression training, is currently not publicly available as its license does not allow for open-source redistribution. However, the pseudocode of all algorithms used in our work, namely, private join and interactive bootstrapping, is provided in *SI Appendix*. Upon request sent to the corresponding author(s), we can provide binaries that, in combination with open-source resources, can be used for the sole purpose of verifying and reproducing the experiments in the manuscript.

Previously published data were used for this work (Our main experiments are based on the public oncological data set described in ref. 29

<https://doi.org/10.1038/s41591-020-0839-y>). The colorectal cancer patients' datasets generated and/or analyzed during the current study are not publicly available due to patients' privacy. Personal patient information was anonymized and stored on a password-protected computer. The computer is located in a locked office of the investigator. The data that support the findings of this study are available from Dr. Ravit Geva but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The data are, however, available from the authors upon reasonable request and with permission of the Tel Aviv Sourasky Medical Center Helsinki Committee.

Author affiliations: ^aTel Aviv Sourasky Medical Center, Tel Aviv 64239, Israel; ^bDana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215; ^cDuality Technologies, Inc., Hoboken, NJ 07103; ^dUniversity of California, San Diego, CA 92093; ^eMassachusetts Institute of Technology, Cambridge, MA 02139; and ^fSimons Institute for the Theory of Computing, University of California, Berkeley, CA 94720

Author contributions: A.G., Y.P., M.B., and S.G. designed research; R.G., A.G., Y.P., L.L., O.R., A.A., N.G., M.B., Z.D., B.W., D.M., F.B., D.T.B., I.W., S.P.-A., T.S., L.A.L., D.M., V.V., and A.A.B. performed research; R.G., A.G., Y.P., L.L., O.R., M.B., Z.D., B.W., D.M., F.B., D.T.B., I.W., S.P.-A., T.S., and L.A.L. analyzed data; and A.G., Y.P., A.A., N.G., M.B., A.A.B., and S.G. wrote the paper.

Reviewers: A.J., Johns Hopkins University; and B.P., Bar Ilan University.

Competing interest statement: S.G. is the director of the Simons Institute at University of California, Berkeley, and is also a cofounder of Duality Technologies to which she consults 1 d a week. V.V. is a Professor at Massachusetts Institute of Technology and is also a cofounder of Duality Technologies to which he consults 1 d a week. D.M. is a Professor at University of California, San Diego, and is also a consultant for Duality Technologies 1 d a week. Y.P., L.L., O.R., A.A., N.G., M.B., Z.D., and A.A.B. are full-time employees of Duality Technologies.

- L. P. Garrison Jr, P. J. Neumann, P. Erickson, D. Marshall, C. D. Mullins, Using real-world data for coverage and payment decisions: The ISPOR real-world data task force report. *Value Health* **10**, 326–335 (2007).
- A. Cave, X. Kurz, P. Arlett, Real-world data for regulatory decision making: Challenges and possible solutions for Europe. *Clin. Pharmacol. Ther.* **106**, 36–39 (2019).
- G. D. Demetri, S. Stacchiotti, Contributions of real-world evidence and real-world data to decision-making in the management of soft tissue sarcomas. *Oncology* **99** (Suppl. 1), 3–7 (2021).
- J. P. Hall *et al.*, Real-world treatment patterns in patients with advanced (stage III-IV) ovarian cancer in the USA and Europe. *Future Oncol.* **16**, 1013–1030 (2020).
- Submitting documents using real-world data and real-world evidence to FDA for drug and biological products (2022). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drug-and-biological-products>. Accessed 20 January 2023.
- P. Naidoo *et al.*, Real-world evidence and product development: Opportunities, challenges and risk mitigation. *Wien Klin. Wochenschr.* **133**, 840–846 (2021).
- L. H. Curtis, J. Brown, R. Platt, Four health data networks illustrate the potential for a shared national multipurpose Big-data network. *Health Aff. (Millwood)* **33**, 1178–1186 (2014).
- S. Adimadhyam *et al.*, Leveraging the capabilities of the FDA's sentinel system to improve kidney care. *J. Am. Soc. Nephrol.* **31**, 2506–2516 (2020).
- M. S. Olson, Can real-world evidence save pharma US\$1 billion per year? A framework for an integrated evidence generation strategy. *J. Comp. Eff. Res.* **9**, 79–82 (2020).
- C. Shore, A. W. Gee, B. Kahn, E. H. Forstg, Eds., "Barriers and disincentives to the use of real-world evidence and real-world data" in *National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Forum on Drug Discovery, Development, and Translation* (National Academies Press, Washington, DC, 2019).
- K. El Emam, S. Rodgers, B. Malin, Anonymising and sharing individual patient data. *BMJ* **350**, h1139 (2015).
- R. Chevrier, V. Foufi, C. Gaudet-Blavignac, A. Robert, C. Lovis, Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *J. Med. Internet Res.* **21**, e13484 (2019).
- T. A. Lasko, S. A. Vinterbo, Spectral anonymization of data. *IEEE Trans. Knowl. Data Eng.* **22**, 437–446 (2010).
- A. C. C. Yao, "How to generate and exchange secrets" in *27th Annual Symposium on Foundations of Computer Science (SFCS 1986)* (1986), pp. 162–167.
- Y. Lindell, Secure multiparty computation. *Commun. ACM* **64**, 86–96 (2020).
- C. Gentry, "Fully homomorphic encryption using ideal lattices" in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, STOC 2009* (Association for Computing Machinery, New York, NY, 2009), pp. 169–178.
- J. H. Cheon *et al.*, *Introduction to Homomorphic Encryption and Schemes* (Springer International Publishing, Cham, 2021), pp. 3–28.
- C. Marcolla *et al.*, Survey on fully homomorphic encryption, theory, and applications. *Proc. IEEE* **110**, 1572–1609 (2022).
- H. Cho, D. J. Wu, B. Berger, Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
- B. Hie, H. Cho, B. Berger, Realizing private and practical pharmacological collaboration. *Science* **362**, 347–350 (2018).
- J. H. Cheon, A. Kim, M. Kim, Y. Song, "Homomorphic encryption for arithmetic of approximate numbers" in *Advances in Cryptology - ASIACRYPT 2017*, T. Takagi, T. Peyrin, Eds. (Springer International Publishing, Cham, 2017), pp. 409–437.
- M. Blatt, A. Gusev, Y. Polyakov, S. Goldwasser, Secure large-scale genome-wide association studies using homomorphic encryption. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11608–11613 (2020).
- A. Kim, Y. Song, M. Kim, K. Lee, J. H. Cheon, Logistic regression model training based on the approximate homomorphic encryption. *BMC Med. Genomics* **11**, 83 (2018).
- K. Han, S. Hong, J. H. Cheon, D. Park, Logistic regression on homomorphic encrypted data at scale. *Proc. AAAI Conf. Artif. Intell.* **33**, 9466–9471 (2019).
- E. Lee *et al.*, Low-complexity deep convolutional neural networks on fully homomorphic encryption using multiplexed parallel convolutions. *PMLR* **162**, 12403–12422 (2022).
- D. Froelicher *et al.*, Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat. Commun.* **12**, 5910 (2021).
- C. Mouchet, J. R. Troncoso-Pastoriza, J. Bossuat, J. Hubaux, Multiparty homomorphic encryption from ring-learning-with-errors. *Proc. Priv. Enhancing Technol.* **2021**, 291–311 (2021).
- G. Asharov *et al.*, "Multiparty computation with low communication, computation and interaction via threshold FHE" in *Advances in Cryptology - EUROCRYPT 2012*, D. Pointcheval, T. Johansson, Eds. (Springer, Heidelberg, 2012), pp. 483–501.
- D. A. Braun *et al.*, Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat. Med.* **26**, 909–918 (2020).
- R. Geva *et al.*, Verification of statistical oncological endpoints on encrypted data: Confirming the feasibility of real-world data sharing without the need to reveal protected patient information. *J. Clin. Oncol.* **39**, e18725 (2021).
- A. A. Badawi, Y. Polyakov, Demystifying bootstrapping in fully homomorphic encryption. *Cryptology ePrint Archive*. Paper 2023/149 (2023). <https://eprint.iacr.org/2023/149>.
- D. Miao *et al.*, Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801–806 (2018).
- S. Groha *et al.*, Germline variants associated with toxicity to immune checkpoint blockade. *Nat. Med.* **28**, 2584–2591 (2022).
- Y. Wu *et al.*, Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, 1891 (2019).
- J. P. Pirruccello *et al.*, Deep learning enables genetic analysis of the human thoracic aorta. *Nat. Genet.* **54**, 40–51 (2022).
- L. T. Elliott *et al.*, Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
- Y. Polyakov, K. Rohloff, G. W. Ryan, D. B. Cousins, PALISADE lattice cryptography library. *Gitlab*. <https://gitlab.com/palisade/palisade-release>. Accessed 20 July 2023.
- Z. Brakerski, Fully homomorphic encryption without modulus switching from classical GapSVP in *Advances in Cryptology - CRYPTO 2012: 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19–23, 2012. Proceedings* (Springer, 2012), pp. 868–886.
- J. Fan, F. Vercauteren, Somewhat practical fully homomorphic encryption. *Cryptology ePrint Archive*. Paper 2012/144 (2012). <https://eprint.iacr.org/2012/144>.
- A. Kim, A. Papadimitriou, Y. Polyakov, "Approximate homomorphic encryption with reduced approximation error" in *Topics in Cryptology - CT-RSA 2022*, S. D. Galbraith, Ed. (Springer International Publishing, Cham, 2022), pp. 120–144.
- A. Kim, Y. Polyakov, V. Zucca, "Revisiting homomorphic encryption schemes for finite fields" in *Advances in Cryptology - ASIACRYPT 2021*, M. Tibouchi, H. Wang, Eds. (Springer International Publishing, Cham, 2021), pp. 608–639.
- M. Ion *et al.*, "On deploying secure computing: Private intersection-sum-with-cardinality" in *IEEE European Symposium on Security and Privacy, EuroSpsampsP2020, Genoa, Italy, September 7–11, 2020* (IEEE, 2020), pp. 370–389.