FedECA: A Federated External Control Arm Method for Causal Inference with Time-To-Event Data in Distributed Settings

Jean Ogier du Terrail

jean.du-terrail@owkin.com

https://orcid.org/0000-0001-6320-819X

Quentin Klopfenstein

Owkin, Inc.

Honghao Li

Owkin, Inc.

Imke Mayer

Owkin, Inc.

Nicolas Loiseau

Owkin Lab, Owkin, Inc.

Mohammad Hallal

Owkin, Inc. https://orcid.org/0000-0003-1624-9647

Michael Debouver

Owkin, Inc.

Thibault Camalon

Owkin, Inc.

Thibault Fouqueray

Owkin, Inc.

Jorge Arellano Castro

Owkin, Inc.

Zahia Yanes

Owkin, Inc.

Laëtitia Dahan

Department of Digestive Oncology, Hôpital la Timone, Marseille, France

Julien Taieb

HEGP hospital Paris descartes university https://orcid.org/0000-0002-9955-4753

Pierre LAURENT-PUIG

Centre de Recherche des Cordeliers https://orcid.org/0000-0001-8475-5459

Jean-Baptiste Bachet

Sorbonne Paris Cité Université and Hôpital Pitié-Salpêtrière, Assistance Publique - Hôpitaux de Paris https://orcid.org/0000-0003-4337-286X

Shulin Zhao

Ruijin Hospital, Shanghai Jiao Tong University School of Medicine https://orcid.org/0000-0001-8233-0115

Rémy Nicolle

INSERM-UMR1149, Centre de Recherche sur l'Inflammation, Paris https://orcid.org/0000-0001-8084-1173

Jérôme Cros

Université Paris Cité - Beaujon Hospital https://orcid.org/0000-0002-4935-3865

Daniel Gonzalez

INSERM UMR1231, Laboratory of Excellence LipSTIC and label Ligue Nationale contre le Cancer, Dijon

Robert Carreras-Torres

Girona Biomedical Research Institute (IDIBGI) https://orcid.org/0000-0002-2925-734X

Adelaida Garcia Velasco

Department of Medical Oncology, Catalan Institute of Oncology, Doctor Josep Trueta University Hospital, Girona

Kawther Abdilleh

Pancreatic Cancer Action Network

Sudheer Doss

Pancreatic Cancer Action Network

Félix Balazard

Owkin, Inc.

Mathieu Andreux

Owkin, Inc.

Article

Keywords:

Posted Date: November 21st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5346692/v1

Additional Declarations: Yes there is potential Competing Interest. The authors declare the existence of a financial competing interest. Some authors are or were employed by Owkin, Inc. during their time on the project (J. O.d.T, Q.K., M.A., H.L, I.M., N.L., M.H., M.D., T.C., T.F.). P. L.-P. has received honoraria for consulting and/or advisory board for AMGEN, Pierre Fabre, Biocartis, Servier and BMS. J.B. Bachet has

received personal fees from Amgen, Bayer, Bristol Myers Squibb, GlaxoSmithKline, Merck Serono, Merck Sharp & Dohme, Pierre Fabre, Sanofi, Servier, and non-financial support from Amgen, Merck Serono, and Roche, outside the submitted work. J. T. has received honoraria as a speaker and/or in an advisory role from AMGEN, Astelllas, Astra Zeneca, Boehringer, BMS, Merck KGaA, MSD, Novartis, ONO pharmaceuticals, Pierre Fabre, Roche Genentech, Sanofi, Servier and Takeda. A. G. V. has received honoraria as a speaker and/or in an advisory role from Astra Zeneca, Merck Serono, MSD, Novartis, Roche Genentech, Sanofi, and Servier. R. N. has received honoraria as a consultant from Cure51. Part of this work corresponding to work-package 4 of the RHU AI-TRIOMPH and carried out by Owkin France was supported by Agence Nationale de la Recherche as part of the France 2030 plan with reference ANR-23-RHUS-0012.

FedECA: A Federated External Control Arm Method for Causal Inference with Time-To-Event Data in Distributed Settings

Jean Ogier du Terrail[*1†], Quentin Klopfenstein[1†], Honghao Li[1†], Imke Mayer[1], Nicolas Loiseau[1], Mohammad Hallal[1], Michael Debouver[1], Thibault Camalon[1], Thibault Fouqueray[1], Jorge Arellano Castro[1], Zahia Yanes[1], Laëtitia Dahan[2], Julien Taïeb[3], Pierre Laurent-Puig[4, 5], Jean-Baptiste Bachet[6], Shulin Zhao[4], Remy Nicolle[7], Jérôme Cros[8], Daniel Gonzalez[9], Robert Carreras-Torres[10], Adelaida Garcia Velasco[11, 10], Kawther Abdilleh[12], Sudheer Doss[12], Félix Balazard[1‡], and Mathieu Andreux[1‡]

[1]Owkin, Inc., New York, NY, USA
[2]Department of Digestive Oncology, Hôpital la Timone, Marseille, France
[3]GI oncology department Georges Pompidou European Hospital, Université Paris Cité, CARPEM CCC, 20 rue leblanc 75015 Paris, APHP, Paris, France
[4]Centre de Recherche des Cordeliers, Sorbonne Université, Inserm, Université Paris Cité, Paris, France
[5]Institut du Cancer Paris CARPEM, AP-HP Centre, Hôpital Européen Georges Pompidou, Paris, France
[6]Sorbonne University, Hepatogastroenterology and digestive oncology department, Pitié Salpêtrière hospital, APHP, Paris, France
[7]Université Paris Cité, Centre de Recherche sur l'Inflammation (CRI), INSERM, U1149, CNRS, ERL 8252, F-75018 Paris, France
[8]Department of Pathology, Université Paris Cité - FHU MOSAIC, Beaujon Hospital, Clichy, France
[9]Fédération Francophone de Cancérologie Digestive, Dijon, France
[10]Institut d'Investigació Biomèdica de Girona (IDIBGI), Girona, Catalonia, Spain
[11]Department of Medical Oncology, Catalan Institute of Oncology, Doctor Josep Trueta University Hospital, Girona, Catalonia, Spain
[12]Pancreatic Cancer Action Network, Manhattan Beach, CA, USA

October 28, 2024

Abstract

External control arms (ECA) can inform the early clinical development of experimental drugs and provide efficacy evidence for regulatory approval. However, the main challenge in implementing ECA lies in accessing real-world or historical clinical trials data. Indeed, regulations protecting patients' rights by strictly controlling data processing make pooling data from multiple sources in a central server often difficult. To address these limitations, we develop a new method, "FedECA" that leverages federated learning (FL) to enable inverse probability of treatment weighting (IPTW) for time-to-event outcomes on separate cohorts without needing to pool data. To showcase the potential of FedECA, we apply it

*Corresponding author
†These authors contributed equally.
‡These authors contributed equally.

in different settings of increasing complexity culminating with a real-world use-case in which FedECA provides evidence for a differential effect between two drugs that would have otherwise gone unnoticed. By sharing our code, we hope FedECA will foster the creation of federated research networks and thus accelerate drug development.

Development of innovative drugs is a long and challenging task with increasing costs [1]. The probability of success of a new drug is low, with about 10% of drugs that enter clinical trials reaching FDA approval [2]. Phase III randomized trials, that aim at establishing clinical efficacy, fail approximately in one case out of two [3]. Improving this rate is crucial to overcome those obstacles and accelerate the access to new drugs while reducing the development costs. Statistical methods were developed to compare the efficacy of a treatment to a control group that is built with data from external sources to the current trial (External Control Arm - ECA). ECA methods account for the potential bias introduced by the non-randomized nature of the control group providing earlier evaluation of efficacy which can inform the transition from a single-arm phase II to a phase III clinical trial [4, 5].

ECA are increasingly used in clinical applications [6] and they are receiving more and more attention from regulatory agencies (FDA, EMA) [7, 8]. Externally controlled trials may also substitute randomized controlled trials (RCT) in specific situations where an RCT would be deemed unfeasible or untimely. This is the case for rare diseases where patient recruitment is difficult and long [9] as well as in some cases in oncology involving specific subgroups of patients [10, 11, 6, 12].

Statistically, the lack of randomization between the treated arm and the external control arm makes a naive comparison between the two susceptible to confounding bias. Therefore, assuming causal identifiability of the treatment effect [13], statistical or machine learning methods [14, 15, 16, 17, 18, 19, 20] are needed to provide valid estimates and inference of the treatment effect in this setting.

Despite progress in statistical methods, a major obstacle to the feasibility of ECA is data sharing. Due to its sensitivity, health data is strictly regulated by, *e.g.*, the general data protection regulation (GDPR) in the EU and the health insurance portability and accountability act (HIPAA) in the US. Even after careful pseudonymization [21] sharing health data remains a complex endeavor notably because of data ownership and liability issues. Thereby, in practice, ECA involving phase II data from a pharmaceutical company and potentially real-world data from different hospitals or medical institutions are difficult to set up.

To address this data sharing challenge, various machine learning techniques were proposed in the past couple of years. Among them, federated learning (FL) [22] a privacy-enhancing technology (PET) allows the extraction of knowledge and training models from multiple institutions without pooling data. It has already been used with success in similar settings to connect both pharmaceutical companies [23] and hospitals [24, 25] in federated research networks.

Herein, we investigate the use of FL to build ECA, focusing on time-to-event outcomes such as progression-free survival (PFS) or overall survival (OS), which are predominant in oncology RCTs [26].

In this work, we introduce FedECA, a federated external control arm method which is a federated version of inverse probability of treatment weighting (IPTW). This method consists of three main steps, each of them performed via FL: training a propensity score model, fitting a weighted time-to-event model, and testing the treatment effect. Thus facilitating ECAs for pharmaceutical companies by giving them access to real-world control patients from multiple institutions while limiting patient data exposure thanks to FL.

We first demonstrate the efficacy of our approach on synthetic data created using realistic data generation processes both with in-RAM simulations and on a deployed federated network with 10 distant synthetic centers located in the cloud.

We show that FedECA reaches identical conclusions to IPTW on pooled data as well as better statistical power compared to a competing method based on federated analytics (FA), while also controlling the moments' differences between the two arms.

Furthermore, we showcase two examples of the application of FedECA to real patients data starting with a simulation of FL using real data from trials before moving to the end-to-end deployment of a real federated research network between three research institutions located in different countries.

1 Results

1.1 FedECA, a federated ECA method

A federated extension of the IPTW method for ECA (FedECA) is deployed. FedECA allows the estimation of a treatment effect by comparing the experimental drug arm, stored in one center, to a control arm defined by external data held within different centers, as illustrated in Figure 1. FedECA uses estimated propensity scores to reweight treated and control patients to balance the two groups and correct for potential biases (see Methods 6.2.1 for more details).

Figure 1 illustrates the advantages of our proposed method, where data can remain on the premises of the participating centers and only aggregated information is shared. This is in contrast to the classical ECA analysis where data is pooled into a single place. Herein, an aggregator node is responsible for the orchestration of the training process, the aggregation and redistribution of the results, without directly seeing raw data.

1.2 FedECA is equivalent to a standard IPTW model trained on pooled data

We demonstrate FedECA's abilities to accurately estimate several key quantities from the IPTW analysis on realistic simulated data. We refer to Methods 6.4.1 for the data generation process details. In this experiment, we compare the results obtained from a classical IPTW analysis where data is pooled into the same place against FedECA operating in distributed settings.

We monitor four key metrics: the propensity scores, the hazard ratio and p-values associated with the treatment allocation covariate, derived from a Wald test, and the partial likelihood of the Cox model. We repeat this simulation 100 times and report the relative error between the pooled IPTW against FedECA in Figure 2. The boxplots represent the distribution of the relative error for each quantity among the 100 repetitions.

The relative errors between FedECA and the pooled IPTW are very small and do not exceed 0.2%, illustrating the effectiveness of the proposed optimization process in a federated fashion. Moreover, Supplementary Figure S1 shows that the number of centers among which the data is split does not impact the performance of FedECA. Hence, we illustrate that up to a negligible error, probably due to finite precision numerical errors in the optimization process, FedECA provides results that are equivalent to the classical IPTW despite not having access to all data in the same location.

1.3 FedECA and MAIC both control SMD

We will compare in the next section the ability of several methods to correctly detect a treatment effect if there is one (statistical power) while not detecting it if there is none (type I error control). The main competitors of FedECA are matching-adjusted indirect comparisons (MAIC) [27], a method that allows to estimate a treatment effect using only aggregated statistics such as the mean and the standard deviation from the treatment arm which are made available to the control arm, and the unweighted version of FedECA (WebDISCO [28]). The reweighting procedure of MAIC involves only communicating statistics a single time between the centers, therefore MAIC's reweighting is a federated analytics method (FA) while FedECA's reweighting uses FL.

Before measuring power and type I error we first assess the performance of the reweighting step of FedECA with respect to the correction of covariate shift which is induced by the non-randomized nature of the two arms and which results in confounding bias in the treatment effect estimation. To do that we estimate the standardized mean difference (SMD) of covariates between the two treatment groups after reweighting. The SMD is a coarse *univariate* measure which summarizes for each covariate the balance between the two groups by only looking at the first and second order moments, see Methods Section 6.3.2. However SMD is expected by regulators [29], which ask ECA methods to control SMD below a threshold (usually 10%). While MAIC explicitly enforces perfect matching of low order moments irrespective of the covariate shift leading to zero SMD *by design* (see Methods Section 6.6), FedECA does *multivariate* balancing through the propensity scores.

We consider scenarios with different levels of covariate shift. In one end of the spectrum, treatment allocation does not depend on the covariates: there is no covariate shift. Therefore the SMDs of all covariates are small even before reweighting. On the other end of the spectrum, treatment allocation depends more and more on the covariates values (details of the data generation are given in Section 6.4.1). Therefore we expect weighting to be necessary to control SMD.

This is illustrated in Figure 3a where we show the mean absolute SMD as a function of the covariate shift for the three different methods: FedECA, MAIC and the non-adjusted treatment effect estimation (unweighted). The covariate shift is a parameter that controls the intensity of the confounding factors on the treatment allocation variable biasing the two groups. A covariate shift of zero is equivalent to a random allocation in the treatment arms (see Methods 6.4.1 for more details). For low covariate shift (≤ 0.5) all three methods control the SMDs of the covariates, while for medium to high covariate shifts (> 0.5), the unweighted method fails to control the SMDs while both MAIC and FedECA control SMD. More details per-covariate are available Figure 3b for the two extremes.

1.4 FedECA outperforms MAIC in power to detect a treatment effect

Figure 3c shows the estimated statistical power and type I error under different varying conditions, including the covariate shift and the number of samples.

One of the key steps influencing the statistical power and type I error is the variance estimation method, applied for each treatment effect point estimation. We compare three variance estimation methods: the bootstrap estimator, the robust sandwich-type estimator, and the naive estimator based on the inversion of the observed Fisher information [16]. For FedECA, only the bootstrap variance estimator successfully controls the type I error at around 5%. In comparison, the robust variance estimator systematically over-estimates the variance, which results in overly-conservative estimations of both the type I errors and the statistical power. This finding is consistent with previous works [16]. Lastly, for FedECA, the naive variance estimator fails to control the type I error. For MAIC, the resampling with replacement during the bootstrap variance estimation can only be done on the individual patient data (IPD) available for the control arm, since in practice the aggregated data from the treatment arm is fixed. Compared to FedECA, it controls the type I error only at small covariate shifts, and loses control when the covariate shift increases. The robust variance estimator shows over-estimations of variance similar to those of FedECA. For comparison, we also consider the unweighted version of FedECA. However, because it cannot account for the confounding effects introduced by the non-randomized nature of the two groups and the resulting covariate shift, it quickly loses control over the type I error as soon as the covariate shift is no longer zero.

When comparing the statistical power of those methods that successfully control the type I error, FedECA with bootstrap variance estimator shows the best performance, followed by FedECA with robust variance estimator. Both variants of FedECA outperform MAIC with the "robust" variance estimator, as the covariate shift and number of samples change.

1.5 FedECA can be used in real-world conditions on synthetic data

We host up to 11 "servers" in the cloud (including an aggregation server) and deploy the Substra [30] software over all centers. Details of the cloud setup are available in Section 6.5. For each experiment we use the first of the servers as the trusted third party performing the aggregation (the "server") and the rest of the servers as data owners holding a different part of the data (the "centers"). Each "center" has a different set of credentials which gives it different permissions over the assets created in the federated network. Each center registers a predefined subset of the synthetic data as if it were its own through the Substra system. One can launch FedECA on the deployed network by simply changing the type of backend used and specifying the identifiers of the datasets (hashes) registered into the Substra platform as inputs to the fit method following `scikit-learn`'s fit API [31]. An example of FedECA python API is given Supplementary Listing S1.

We monitor the runtime of the full pipeline when it runs in-RAM and over the cloud as a function of the number of centers. We give conservative estimates by setting a target number of rounds for the training of the propensity model and the Cox model to be high (20) as well as compute the federated robust sandwich estimator, which adds overhead and is not necessary if using the bootstrap variance estimator.

Experiment (Treatment vs. Control)	Method	Treatment data source	Control data source	HR (95% CI)	p
Apa-AA-P vs. AA-P	FedECA (bootstrap)	Trial 1	Trial 2	0.66 (0.55, 0.78)	**<0.00001**
	Literature [34]	Trial 1	Trial 1	0.70 (0.60-0.83)	**<0.0001**
AA-P vs. P	FedECA (bootstrap)	Trial 1	Trial 2	0.50 (0.42, 0.59)	**<0.000001**
	Literature [35]	Trial 2	Trial 2	0.53 (0.45-0.62)	**<0.001**
AA-P vs. AA-P	FedECA (bootstrap)	Trial 1	Trial 2	0.90 (0.77, 1.05)	0.18
Apa-AA-P vs. P	FedECA (bootstrap)	Trial 1	Trial 2	0.37 (0.31, 0.45)	**<0.000001**

Table 1: Estimation of treatment effect on radiographic progression-free survival by comparing different regimens across different trials. Trial 1 refers to NCT02257736, Trial 2 refers to NCT00887198.

In-RAM experiments take a few seconds with 10 centers which is to be compared with IPTW on pooled data which has a below-second runtime. This slowdown is due mostly to (1) processing each client sequentially (2) the static nature of the Substra framework which forces one to run a target number of rounds higher than needed for convergence (see Section 6.7.2 for more explanations). While (2) is a fundamental limitation of Substra, (1) could be improved upon by using Python multiprocessing.

The real-world runtime is almost constant with respect to the number of clients and takes just under 2 hours (1h18min on average with a standard deviation of approximately 3 minutes). We provide in Table Supplementary Table S1 the full breakdown of the different runtimes across settings.

Insofar as 10 centers is already large in the considered cross-silo FL setup, this result hints at a good scalability in terms of speed, provided an appropriate infrastructure can be deployed across the different centers. Our result is consistent with previous Substra deployments [23].

1.6 FedECA can be used on real world use-cases

1.6.1 Application on real prostate cancer data with simulated federated learning

We access data of two phase III metastatic castration-resistant prostate cancer trials from the Yale University Open Data Access (YODA) project [32, 33]. In all cases we simulate a FL setup in which the data is held by two *synthetic* centers, one that holds the treated arm and the other the rest of the patients. We note that according to our experiments Supplementary Figure S1, the number of centers does not impact the performance of FedECA (nor should how the data is distributed across the different centers), therefore we expect roughly the same results if we had chosen different data splits. All information about cohort construction is available in Methods 6.4.2.

In the following, we present results focusing on average treatment effect inference on the radiographic progression-free survival (rPFS) endpoint comparing regimens across different trials. We show first the results reported for each trial found in the associated publications. Then, we conduct simulational ECA studies by replacing the abiraterone acetate + prednisone (AA-P) arm in each trial with the same arm from the other trial; the results of which can be considered as external negative control validations. Additionally, we compare the two AA-P arms as a way to detect the potential confounding between the two study populations, as well as to validate the previous ECA analyses. Finally, we compare the apatulamide + abiraterone acetate + prednisone (Apa-AA-P) arm with the prednisone (P) arm, which is not seen in the literature, and which showcases the potential usefulness of our method to provide new evidence which is otherwise unavailable or time and resource expensive to produce.

In Table 1, we show consistency of the estimations of treatment effect when compared with the published results, as well as the non-significant treatment effect when comparing two AA-P arms. In addition, we show a strong treatment effect of the regimen Apa-AA-P versus P, which is reasonable considering the superiority of Apa-AA-P versus AA-P, and of AA-P versus P.

Method	#centers	HR (95% CI)	z	p
FedECA (bootstrap)	FFCD, IDIBGI, PanCAN	0.75 (0.62, 0.92)	−2.80	0.005 05
FedECA (robust)	FFCD, IDIBGI, PanCAN	0.75 (0.61, 0.92)	−2.74	0.0061
FedECA (naïve)	FFCD, IDIBGI, PanCAN	0.75 (0.66, 0.86)	−4.10	0.000 041 0
IPTW (bootstrap)	FFCD	1.13 (0.80, 1.61)	0.71	0.480
IPTW (bootstrap)	IDIBGI	0.74 (0.40, 1.37)	−0.95	0.342
IPTW (bootstrap)	PanCAN	0.80 (0.55, 1.16)	−1.19	0.234

Table 2: Effect on overall survival of FOLFIRINOX over gemcitabine + nab placlitaxel as estimated by FedECA versus single-center estimates.

1.6.2 Application on real metastatic pancreatic cancer data in a real deployed federated research network

We deploy a federated network across three cancer centers: the Fédération Francophone de Cancérologie Digestive (FFCD) holding the data from two past clinical trials data and the Institut d'Investigació Biomèdica de Girona (IDIBGI) and the Pancreatic Cancer Action Network (PanCAN) that hold data from clinical practice totalling $n = 514$ patients. All information about cohort construction is available in Methods 6.4.3. We compare the efficiency of FOLFIRINOX over gemcitabine and nab-paclitaxel on overall survival. Data from each of the centers in the network is described in Methods (see Table Supplementary Table S7, Supplementary Table S5 and Supplementary Table S6 for more details).

Figure 4a demonstrates that the propensity model in FedECA, trained with FL, effectively balances the two patient groups, reducing the standardized mean difference (SMD) between the two arms to below 10%, which was not achieved before reweighting.

Table 2 shows evidence for an effect on overall survival of FOLFIRINOX over gemcitabine and nab-paclitaxel with a hazard ratio (HR) of $0.75\,(\text{IC} = 0.63, 0.92)$ and an associated p-value of 0.005045, which is exactly in line with the literature (HR 0.77) [36] with IPTW on pooled data. Single-center analyses on each center presented in Table 2 are underpowered to detect this effect. Figure 4 shows the results of this real-world use-case in terms of the estimated effect and illustrated by propensity-weighted Kaplan-Meier curves. The propensity-weighted Kaplan-Meier estimator of the full cohort as well as the global SMD represented respectively in Figure 4b and Figure 4a were obtained through the application of federated analytics without pooling data, see Sec 6.3.

2 Discussion

In spite of the constraints on the distribution of the computations, we have shown that FedECA accurately replicates its pooled-equivalent counterpart IPTW up to machine precision, see Figure 2. This directly imbues FedECA with the same statistical properties as the well-established IPTW method.

Indeed, FedECA compares favorably to a simpler federated analytic baseline MAIC in terms of statistical power and type I error while controlling for standardized mean difference (below 10%) even in the presence of strong covariate shifts, see Figure 3c.

In contrast with its numerous stratified competitors [37, 38, 39, 40, 41], FedECA can be applied to the setting of interest, namely the non-stratified setting where all treated patients are held by a pharmaceutical company and all control patients are disseminated across multiple different institutions, i.e., treatment and control patients are in separate locations. FedECA also remains applicable in the general case where treated patients spread across different centers, mixed or not with control patients as we notably demonstrate in the metastatic pancreatic adenocarcinoma use-case. We believe that in the case of drug development this setting of building an external control arm from several other centers is the most relevant but other cases with real-world data could be considered.

An additional advantage of FedECA is that it has the same flexibility as IPTW in terms of the quantities it can estimate. MAIC does not yield the same estimands as IPTW in the context of ECA [11]. More specifically, while MAIC allows only to estimate the average treatment effect on the control (ATC) without

any additional assumptions, IPTW, and by extension FedECA, can be used to estimate the ATE, the average treatment effect on the treated (ATT), as well as the ATC.

Furthermore, FedECA, as a federated extension of IPTW, also enables covariate adjustment through *adjusted* IPTW, which might give researchers more flexibility over which effect measure to estimate whether it is marginal or conditional [42] as both are different and generally require different estimation approaches in the context of time-to-event outcomes due to the non-collapsibility of the hazard ratio.

We show that FedECA is not only a methodology but also a software that can be deployed in real-world settings with different degrees of complexity. In a first experiment we show that with FedECA, we can not only reproduce results of each trial [34, 35, 43], but that FedECA also enables additional and novel analyses that are not feasible using each of the trials separately. While data used in this experiment are not physically distributed in different locations, the results of those simulations are representative of what could be performed on similar data in a real federated setting.

Finally in a second experiment we deployed a Substra federated research network gathering three cancer centers: the fédération française du colon (FFCD, France), the institut d'investigació biomèdica de Girona (IDIBGI, Spain) and the Pancreatic Cancer Action Network (PanCAN, USA) to measure the ATE of FOLFIRINOX (leucovorin and fluorouracil plus irinotecan and oxaliplatin) over gemcitabine and nab-paclitaxel [44, 45, 46, 47] managing to reproduce the results from meta and pooled analyses from the literature in this more complex privacy-enhanced federated setting.

We show that FedECA is the only method in this setting that is able to show evidence of an effect of FOLFIRINOX over gemcitabine and nab-paclitaxel on overall survival due to the added statistical power brought by the federation, as per-center analyses are underpowered due to their limited number of patients as shown in Figure 4..

Moreover, the HR 0.75 measured by FedECA matches almost exactly the one reported in [36]. While the results of our analysis rely on a smaller sample size ($n = 514$) than the largest previous efforts [48, 36, 49, 50, 47], it brings additional evidence on this topic of rising medical interest [51].

We now explore the different limitations of FedECA as well as potential extensions.

We and others [36, 34, 35, 43] focus on the (log) hazard ratio as the effect measure. Other effect measures have been proposed for time-to-event outcomes such as contrasts of restricted mean survival time (RMST) [52]. The latter has the advantage of being collapsible and, as argued in certain applications, of offering better clinical relevance and interpretability [53]. An IPTW-based estimator for the difference of RMST as an effect measure has been proposed [54] and could guide an extension of FedECA to RMST-based effect estimation in a federated setting.

As in WebDISCO [28], Breslow approximation of the Cox log-likelihood is used with respect to ties. This approximation is standard in survival analysis and is currently the default of the widely used `sksurv` package [55]. However this approximation was shown to be biased when the number of ties grows [56]. In Supplementary Figure S3, we provide some experiments artificially increasing the number of ties in the data and study validity with respect to `lifelines`'s [57] implementation that uses the Efron approximation [58]. For a realistic number of ties, *e.g.*, less than 10% in total, FedECA gives valid estimations but performance degrades sharply over this threshold. We note that Efron's approximation, which is more precise, adds complexity to the federation and would need to be investigated by future works.

IPTW was chosen as the main bias correcting method for federating ECA mainly because of its strong performance in settings with small sample sizes compared to propensity score matching methods [14].

As for all causal inference methods, defining the correct set of confounders is a key element of an ECA based on propensity scores [59]. FedECA remains sensitive to misspecification in the propensity score method as its pooled version, IPTW. When building an ECA, one should carefully select the confounders to include in the propensity score method and should consider the possibility of unmeasured confounders as well as ways to perform sensitivity analysis to assess the robustness of the results [60]. On top of that, when performing an ECA analysis, one should ensure that the variables are collected similarly across centers and more importantly that the endpoints of interest are defined similarly across centers. This is a key element to ensure the validity of the methods and we refer to the progression free survival (PFS) endpoint as an example where the definition of the event is not standardized and can lead to different results.

Real-world data is by nature very noisy and does contain missing values or missing features as was the case in the metastatic pancreatic adenocarcinoma use case we considered. While there is a whole literature on missing data imputation in machine learning and the impact of missing value on causal inference analyses

with recent developments being implemented [61, 62, 63, 64, 65, 66] and benchmarked on health data [67], this problem has yet to be tackled satisfactorily in federated settings. In this study, we use a naive solution, which is to apply one of these techniques, MissForest [62], per-site which is suboptimal and could possibly further increase heterogeneity and biases.

Federation of both the propensity model and the Cox proportional hazard (PH) model involves communicating aggregated information to the server and all information transmitted can theoretically be exploited by attackers if they were not generated with local differential privacy (DP).

Notably the federation of the Cox PH model on top involves communicating propensity scores aggregated over per-client risk sets. Such risk sets might, for some actors, be considered sensitive as one could again infer information about IPD by observing sums of scalar products of IDP with the differentially private propensity model weights but we are not aware of any work that would have performed such an attack (see Methods Section 6.2.7).

We tested the application of differential privacy to the first part of the FedECA training, which is the training of the propensity model. However, it showed already to be detrimental to the statistical analysis (see Supplementary Figure S2).

We note that if time-to-event outcomes are considered non-sensitive in our work, as is common in RCTs with the publication of Kaplan-Meier curves [68], discretization and randomization of time-to-event buckets could be considered as an extension of this work to more privacy stringent settings. However, as with all privacy-enhancing layers built on noise addition, this could potentially have a high negative impact to the downstream statistical analysis and would require an even larger pool of patients to remain reliable.

It is an interesting question to ask whether the use of DP in the clinical trials of the future is warranted as it trades off accuracy of treatment effect estimation for data privacy. We do not pretend answering this question in this work.

Another privacy enhancing layer that could be added to FedECA would be to use secure aggregation (SA) [69] to hide individual contributions through cryptographic operations. This would provably hide per-client risk sets and would also allow private set unions (PSU) [70] to compute the global event times or can also be used to secure the aggregator node. Even if quantization due to the use of integers would probably impact the accuracy of the method, it has been shown that with sufficient precision, SA could be a suitable alternative for similar optimization schemes [69, 71].

In this work we presented FedECA, a federated method to perform ECA analysis in a federated setting where treated patient data are in a distinct center and the external control arm is split across different centers that cannot share their data. FedECA is a federated extension of IPTW that reproduces the result of a pooled analysis, yielding similar treatment effect estimation with similar statistical guarantees. FedECA is a suitable method to perform causal inference in distributed ECA settings while limiting IPD exposure. We demonstrated that FedECA is not only a simulation tool but that it can be applied to real use-cases showcasing its ability in two clinically different contexts.

Implementing federated methods in real-world environments can be challenging. This implementation is based on an open source FL software hosted by the LFAI, that has already successfully been used in the targeted high-security healthcare setting with both pharmaceutical companies and cancer centers [25, 23] and that was again validated by the IT teams of each of the participating cancer centers in the metastatic pancreatic adenocarcinoma experiment.

We believe that this implementation will help the adoption of FedECA and will facilitate the development of partnerships across hospitals and medical centers to compare treatment effects in real-world settings.

3 Acknowledgements

This study, carried out under YODA Project 2023-5198, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C..

The data coming from the Pancreatic Cancer Action Network was collected through the SPARK health data platform [72] and derives from the "PanCAN Know Your Tumor Program". We especially thank the

patients that shared their data to this program.

Part of this work corresponding to work-package 4 of the RHU AI-TRIOMPH and carried out by Owkin France was supported by Agence Nationale de la Recherche as part of the France 2030 plan with reference ANR-23-RHUS-0012.

We thank Cameron Davidson Pilon for developing the wonderful python package that is `lifelines` and that was a source of inspiration for this work. We thank Arthur Pignet for discussions on efficient bootstrap implementation with substra as well as for his code allowing efficient distributed computation of moments. We thank Constance Beguier for her help in deriving the distributed computation of the Kaplan-Meier estimator. We thank Maylis Largeteau and Parjeet Kaur for their help in writing the FedECA license. Finally, we thank Jean-Philippe Vert and Nathan Noiry for their insightful comments and suggestions.

4 Author Contributions Statement

M.A. and F.B. conceived the idea of investigating federated methods for external control arms and supervised the paper writing. J. O.d.T., Q.K. and H.L. wrote the paper and led the research. J. O.d.T. designed all the federated algorithms. Q.K. implemented the data generation process as well as the pooled baselines with the help of H.L. and J. O.d.T implemented all federated learning code. M.A. and J. O.d.T wrote the differential privacy federated code together. H.L. and I.M. helped writing the paper and performing experiments. Notably H.L. performed the YODA experiments with the help of I.M. and J.O.d.T. N. L. reviewed the statistical methodology and helped writing the paper. M. H. provided support for creating the figures and feedbacks on the writing of the paper. M. D. on one side and T. C. and T. F. on the other were in charge of respectively administrating the federated learning network infrastructure and deploying Substra. L. D., J. T., P. L.-P., J.-B. B., S. Z., R. N. and J. C. were responsible for the data from the PRODIGE35 and PRODIGE37 trials held at FFCD. D. G. reviewed the paper. In addition to providing data from listed clinical trials, J. T. discussed the study design and reviewed the manuscript

R. C. T. and A. G. V. curated the data from IDIBGI, K. A., S. D. and A. B. curated the data from PanCAN under the guidance of J. A. C. and Z. Y. in charge of data-harmonization and specifications.

5 Competing Interests

The authors declare the existence of a financial competing interest.

Some authors are or were employed by Owkin, Inc. during their time on the project (J. O.d.T, Q.K., M.A., H.L, I.M., N.L., M.H., M.D., T.C., T.F.).

P. L.-P. has received honoraria for consulting and/or advisory board for AMGEN, Pierre Fabre, Biocartis, Servier and BMS.

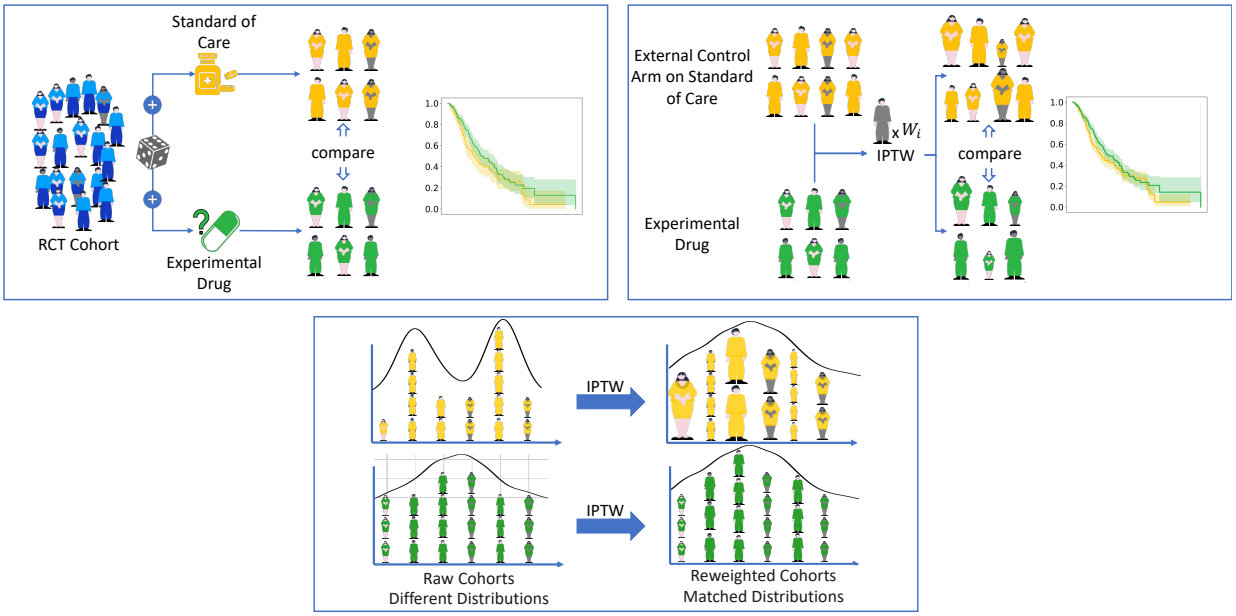
J.B. Bachet has received personal fees from Amgen, Bayer, Bristol Myers Squibb, GlaxoSmithKline, Merck Serono, Merck Sharp & Dohme, Pierre Fabre, Sanofi, Servier, and non-financial support from Amgen, Merck Serono, and Roche, outside the submitted work.

J. T. has received honoraria as a speaker and/or in an advisory role from AMGEN, Astelllas, Astra Zeneca, Boehringer, BMS, Merck KGaA, MSD, Novartis, ONO pharmaceuticals, Pierre Fabre, Roche Genentech, Sanofi, Servier and Takeda.

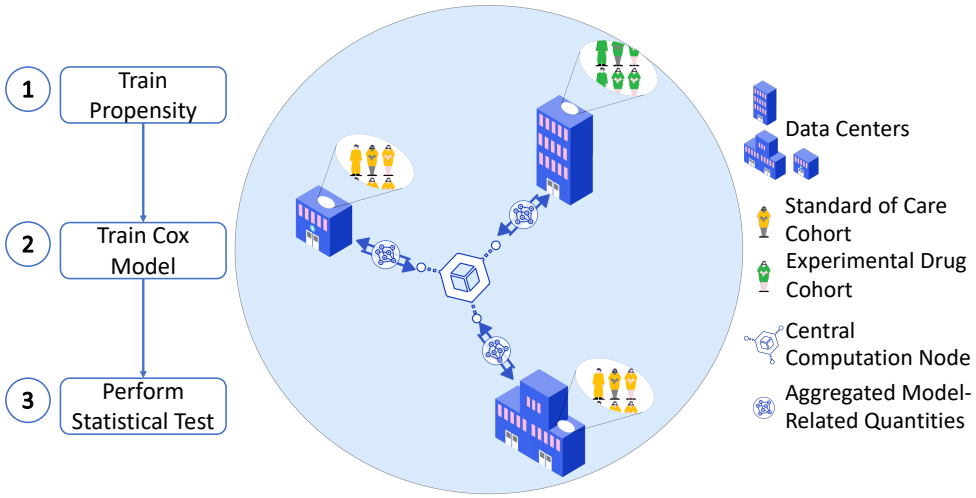
A. G. V. has received honoraria as a speaker and/or in an advisory role from Astra Zeneca, Merck Serono, MSD, Novartis, Roche Genentech, Sanofi, and Servier.

R. N. has received honoraria as a consultant from Cure51.

Part of this work corresponding to work-package 4 of the RHU AI-TRIOMPH and carried out by Owkin France was supported by Agence Nationale de la Recherche as part of the France 2030 plan with reference ANR-23-RHUS-0012.



(a) Illustration of randomized controlled trials (RCT) versus an external control arm (ECA) analysis.



(b) Federated ECA setup.

Figure 1. FedECA graphical abstract. (a) In an RCT, patients are randomly assigned to either the experimental (i.e. treatment) or the control arm. In an ECA, patients are assigned to the treatment arm, while the control arm is defined using historical data. Due to this absence of randomization and the resulting confounding, the two groups of patients cannot be compared directly. To overcome this issue, a model is used to capture the association between the treatment allocation and the confounding factors. From this model, weights are computed and are used to balance the two arms to ensure comparability. Then, the weights are incorporated into a Cox model to estimate the treatment effect. Finally a statistical test is performed to assess the significance of the measured treatment effect. (b) In the considered setting, patient data is stored in different geographically distinct centers and a similar analysis as in (a) is attempted thanks to our algorithm FedECA. A trusted third party is responsible for the orchestration of the training processes, which consists of exchanging model related quantities across the centers. No individual patient data is shared between the centers and only aggregated information is exchanged, which limits patient data exposure while producing equivalent results.

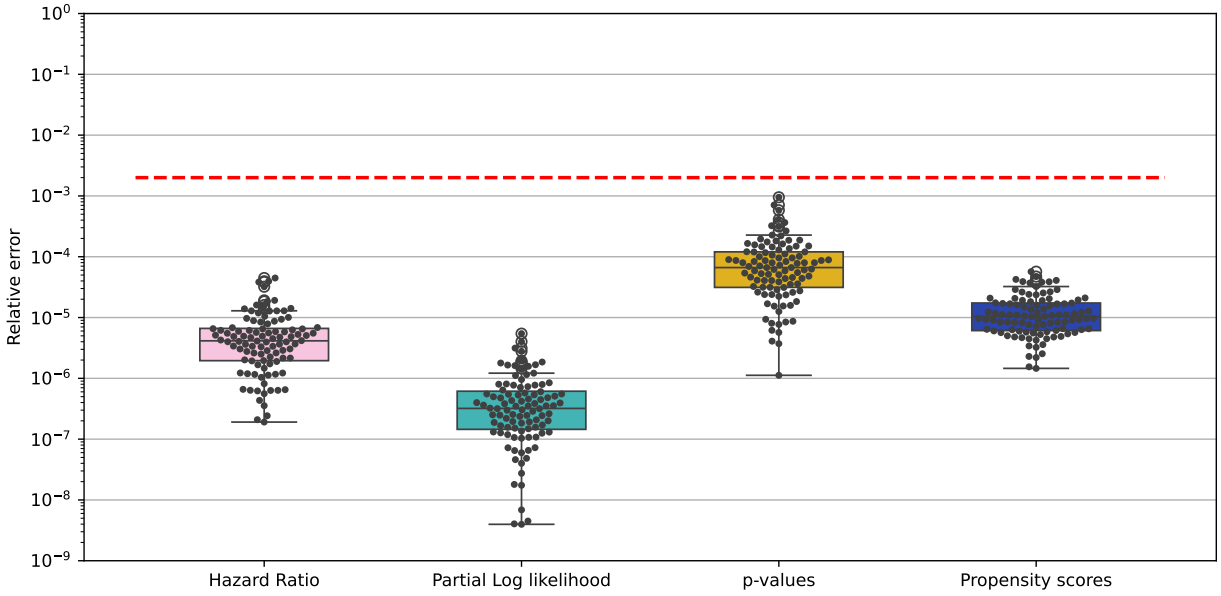
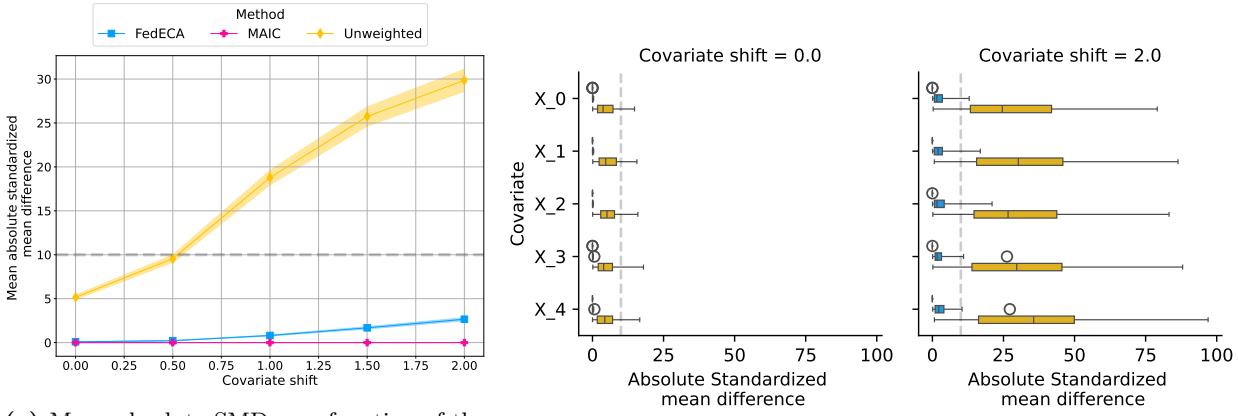
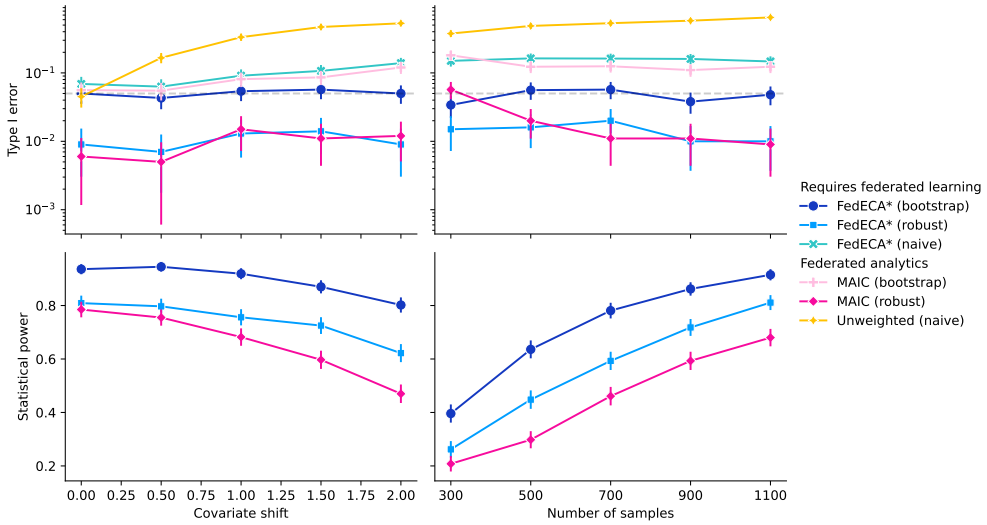


Figure 2. Pooled equivalence between IPTW and FedECA. Box- and swarm-plots of the relative error between the pooled IPTW and the FedECA algorithm on four different quantities: the propensity scores estimated from the logistic regression, the hazard ratio (representing the treatment effect), the p-values associated to the treatment allocation variable (Wald test) and the partial likelihood resulting from the Cox model. The errors were computed on simulated data with 100 repetitions. The red dotted line represents a relative error of 0.2% between the pooled IPTW and FedECA.



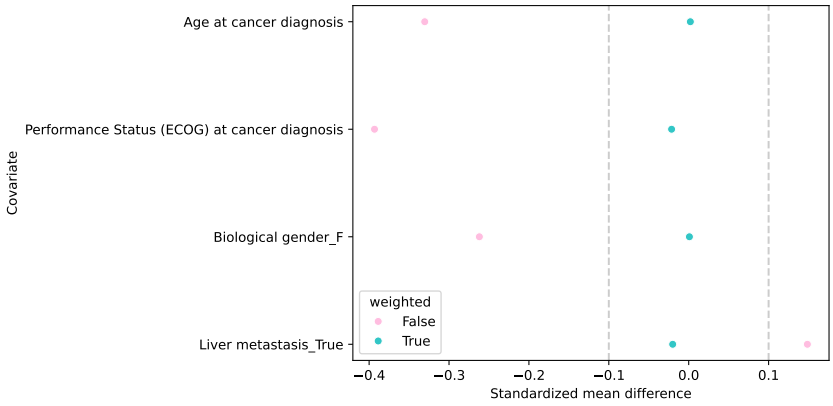
(a) Mean absolute SMD as a function of the covariate shirt.

(b) Absolute SMD for each covariate.

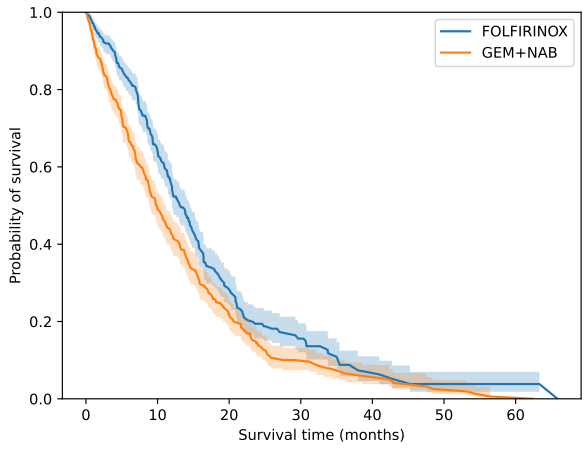


(c) Comparison of different methods on statistical power and type I error of treatment effect estimation.

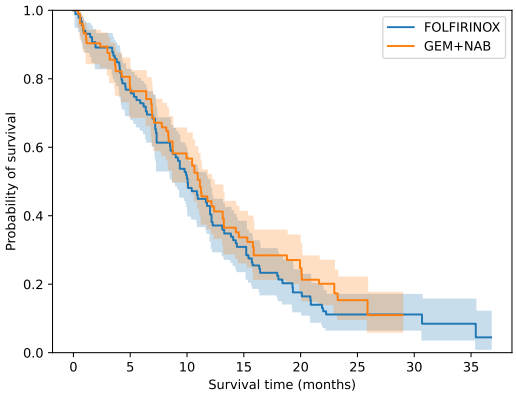
Figure 3. Comparison of different methods on statistical power, type I error of treatment effect estimates as well as standardized mean difference (SMD) of covariates between the two treatment arms. (a) Curves representing the mean absolute SMD computed on 10 covariates as a function of the covariate shift for three different methods: FedECA, MAIC and the non-adjusted treatment effect estimation (unweighted). (b) Boxplots representing the distribution of the absolute SMD over the 100 repetitions for the first five covariates. Each estimation of SMD is based on 100 repetitions of propensity score estimation. For all simulations, we generate 10 covariates and 1000 samples. (c) Different variance estimation methods leading to different p-values are given in parentheses after each method giving point estimates of the hazard ratio. In particular, the naive variance estimation is based on the simple inversion of the observed Fisher information. For statistical power, only results of methods that consistently control the type I error around/under 0.05 (marked by grey dashed lines in top panels) are shown. Each estimation of statistical power or type I error is based on 1000 repetitions of treatment effect estimation. For bootstrap-based variance estimating methods, the number of bootstrap resampling is set to 200. For all simulations, we assume 10 covariates. The hazard ratio of the simulated treatment effect is set to 0.4 for the estimation of statistical power, and to 1.0 for the estimation of type I error. For simulations with varying covariate shifts (the two panels on the left), the number of samples is fixed at 700. For simulations with varying sample size (the two panels on the right), the covariate shift is fixed at 2.0. The asterisk on FedECA indicates that, due to the time-consuming nature of the power analysis, their more lightweight pooled-equivalent counterparts were used instead (pooled IPTW) (see Section 1.2). For confidence intervals we use the central limit theorem applied to Bernoulli variables.



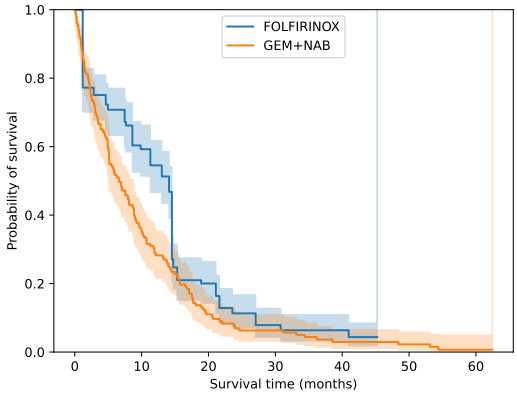
(a) SMD for each covariate before and after weighting.



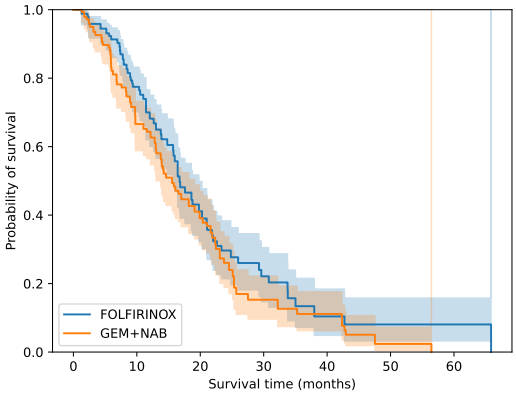
(b) Federated weighted Kaplan-Meier estimator of the full-cohort



(c) weighted Kaplan-Meier estimator of FFCD



(d) weighted Kaplan-Meier estimator of IDIBGI.



(e) weighted Kaplan-Meier estimator of PanCAN.

Figure 4. Real-world FOLFIRINOX effect estimation using FedECA versus local analyses. (a) SMD for each covariate between the two arms of the full cohort before and after weighting by the FedECA propensity model. (b) weighted-Kaplan-Meier curves of the full cohort using FedECA's propensity model. (c) weighted-Kaplan-Meier curve obtained by applying IPTW locally in FFCD using only FFCD held data. (d) weighted-Kaplan-Meier curve obtained by applying IPTW locally in IDIBGI using only IDIBGI held data. (e) weighted-Kaplan-Meier curve obtained by applying IPTW locally in PanCAN using only PanCAN held data.

References

[1] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. Journal of health economics, 47:20–33, 2016.

[2] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. Nature biotechnology, 32(1):40–51, 2014.

[3] Helen Dowden and Jamie Munro. Trends in clinical success rates and therapeutic focus. Nat. Rev. Drug Discov, 18(7):495–6, 2019.

[4] Steffen Ventz, Albert Lai, Timothy F Cloughesy, Patrick Y Wen, Lorenzo Trippa, and Brian M Alexander. Design and evaluation of an external control arm using prior clinical trials and real-world data. Clinical Cancer Research, 25(16):4993–5001, 2019.

[5] Xiang Yin, Ruthanna Davi, Elizabeth B Lamont, Premal H Thaker, William H Bradley, Charles A Leath III, Kathleen M Moore, Khursheed Anwer, Lauren Musso, and Nicholas Borys. Historic clinical trial external control arm provides actionable gen-1 efficacy estimate before a randomized trial. JCO Clinical Cancer Informatics, 7:e2200103, 2023.

[6] Xiaomeng Wang, Flavio Dormont, Christelle Lorenzato, Aurélien Latouche, Ramon Hernandez, and Roman Rouzier. Current perspectives for external control arms in oncology clinical trials: Analysis of ema approvals 2016-2021. Journal of Cancer Policy, page 100403, 2023.

[7] Center for Drug Evaluation, Center for Biologics Evaluation Research, and Oncology Center of Excellence Research. Considerations for the design and conduct of externally controlled trials for drug and biological products. `https://www.fda.gov/media/164960/download`, 2023.

[8] European Medicines Agency. Reflection paper on establishing efficacy based on single arm trials submitted as pivotal evidence in a marketing authorisation. `https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing_en.pdf`, 2023.

[9] Artak Khachatryan, Stephanie H Read, and Terri Madison. External control arms for rare diseases: building a body of supporting evidence. Journal of Pharmacokinetics and Pharmacodynamics, pages 1–6, 2023.

[10] P.S. Mishra-Kalyani, L. Amiri Kordestani, D.R. Rivera, H. Singh, A. Ibrahim, R.A. DeClaro, Y. Shen, S. Tang, R. Sridhara, P.G. Kluetz, J. Concato, R. Pazdur, and J.A. Beaver. External control arms in oncology: current use and future directions. Annals of Oncology, 33(4):376–383, 2022.

[11] Jérôme Lambert, Etienne Lengliné, Raphaël Porcher, Rodolphe Thiébaut, Sarah Zohar, and Sylvie Chevret. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. Blood advances, pages bloodadvances–2022009167, 2022.

[12] Donna Przepiorka, Chia-Wen Ko, Albert Deisseroth, Carolyn L Yancey, Reyes Candau-Chacon, Haw-Jyh Chiu, Brenda J Gehrke, Candace Gomez-Broughton, Robert C Kane, Susan Kirshner, et al. Fda approval: blinatumomab. Clinical Cancer Research, 21(18):4035–4039, 2015.

[13] James M Robins. Data, design, and background knowledge in etiologic inference. Epidemiology, pages 313–320, 2001.

[14] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424, 2011.

[15] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19):2937–2960, 2004.

[16] Peter C Austin. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Statistics in medicine, 35(30):5642–5655, 2016.

[17] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical modelling, 7(9-12):1393–1512, 1986.

[18] Arthur Chatton, Florent Le Borgne, Clémence Leyrat, Florence Gillaizeau, Chloé Rousseau, Laetitia Barbin, David Laplaud, Maxime Léger, Bruno Giraudeau, and Yohann Foucher. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Scientific reports, 10(1):9219, 2020.

[19] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[20] Nicolas Loiseau, Paul Trichelair, Maxime He, Mathieu Andreux, Mikhail Zaslavskiy, Gilles Wainrib, and Michael GB Blum. External control arm analysis: an evaluation of propensity score approaches, g-computation, and doubly debiased machine learning. BMC Medical Research Methodology, 22(1):1–13, 2022.

[21] Christian Ohmann, Rita Banzi, Steve Canham, Serena Battaglia, Mihaela Matei, Christopher Ariyo, Lauren Becnel, Barbara Bierer, Sarion Bowers, Luca Clivio, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ open, 7(12):e018647, 2017.

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017.

[23] Martijn Oldenhof, Gergely Ács, Balázs Pejó, Ansgar Schuffenhauer, Nicholas Holway, Noé Sturm, Arne Dieckmann, Oliver Fortmeier, Eric Boniface, Clément Mayer, Arnaud Gohier, Peter Schmidtke, Ritsuya Niwayama, Dieter Kopecky, Lewis Mervin, Prakash Chandra Rathi, Lukas Friedrich, András Formanek, Peter Antal, Jordon Rahaman, Adam Zalewski, Wouter Heyndrickx, Ezron Oluoch, Manuel Stößel, Michal Vančo, David Endico, Fabien Gelus, Thaïs de Boisfossé, Adrien Darbier, Ashley Nicollet, Matthieu Blottière, Maria Telenczuk, Van Tien Nguyen, Thibaud Martinez, Camille Boillet, Kelvin Moutet, Alexandre Picosson, Aurélien Gasser, Inal Djafar, Antoine Simon, Ádám Arany, Jaak Simm, Yves Moreau, Ola Engkvist, Hugo Ceulemans, Camille Marini, and Mathieu Galtier. Industry-scale orchestrated federated learning for drug discovery. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[24] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. Nature communications, 13(1):7346, 2022.

[25] Jean Ogier du Terrail, Armand Leopold, Clément Joly, Constance Béguier, Mathieu Andreux, Charles Maussion, Benoît Schmauch, Eric W Tramel, Etienne Bendjebbar, Mikhail Zaslavskiy, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature medicine, 29(1):135–146, 2023.

[26] Jennifer Le-Rademacher and Xiaofei Wang. Time-to-event data: an overview and analysis considerations. Journal of Thoracic Oncology, 16(7):1067–1074, 2021.

[27] James E. Signorovitch, Vanja Sikirica, M. Haim Erder, Jipan Xie, Mei Lu, Paul S. Hodgkins, Keith A. Betts, and Eric Q. Wu. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. Value in Health, 15(6):940–947, 2012.

[28] Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. WebDISCO: a web service for distributed Cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association, 22(6):1212–1219, 2015.

[29] US Food, Drug Administration, et al. Considerations for the design and conduct of externally controlled trials for drug and biological products. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products`, February 2023.

[30] Mathieu N Galtier and Camille Marini. Substra: a framework for privacy-preserving, traceable and collaborative machine learning. arXiv preprint arXiv:1910.11567, 2019.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

[32] Harlan M Krumholz and Joanne Waldstreicher. The yale open data access (yoda) project–a mechanism for data sharing. The New England journal of medicine, 375(5):403–405, 2016.

[33] Joseph S Ross, Joanne Waldstreicher, Stephen Bamford, Jesse A Berlin, Karla Childers, Nihar R Desai, Ginger Gamble, Cary P Gross, Richard Kuntz, Richard Lehman, et al. Overview and experience of the yoda project with clinical trial data sharing after 5 years. Scientific data, 5(1):1–14, 2018.

[34] F. Saad, E. Efstathiou, G. Attard, T. W. Flaig, F. Franke, O. B. Goodman, S. Oudard, T. Steuber, H. Suzuki, D. Wu, K. Yeruva, P. De Porre, S. Brookman-May, S. Li, J. Li, S. Thomas, K. B. Bevans, S. D. Mundle, S. A. McCarthy, and D. E. Rathkopf. Apalutamide plus abiraterone acetate and prednisone versus placebo plus abiraterone and prednisone in metastatic, castration-resistant prostate cancer (ACIS): a randomised, placebo-controlled, double-blind, multinational, phase 3 study. Lancet Oncol, 22(11):1541–1559, Nov 2021.

[35] C. J. Ryan, M. R. Smith, J. S. de Bono, A. Molina, C. J. Logothetis, P. de Souza, K. Fizazi, P. Mainwaring, J. M. Piulats, S. Ng, J. Carles, P. F. Mulders, E. Basch, E. J. Small, F. Saad, D. Schrijvers, H. Van Poppel, S. D. Mukherjee, H. Suttmann, W. R. Gerritsen, T. W. Flaig, D. J. George, E. Y. Yu, E. Efstathiou, A. Pantuck, E. Winquist, C. S. Higano, M. E. Taplin, Y. Park, T. Kheoh, T. Griffin, H. I. Scher, D. E. Rathkopf, A. Boyce, A. Costello, I. Davis, V. Ganju, L. Horvath, R. Lynch, G. Marx, F. Parnis, J. Shapiro, N. Singhal, M. Slancar, G. Van Hazel, S. Wong, D. Yip, P. Carpentier, D. Luyten, S. Rottey, F. Van Aelst, T. Cheng, J. Chin, S. Ellard, Y. Fradet, M. Gleave, A. Joshua, L. Klotz, H. Martins, S. North, S. Abdel-Hamid, M. Colombel, A. chon, O. Haillot, F. Joly, S. Oudard, F. Priou, E. Raymond, P. Albers, M. Boegemann, J. Gleissner, J. Gschwend, P. Hammerer, A. Heidenreich, M. Kuczyk, K. Miller, R. Oetzel, J. Roigas, T. Steuber, M. ckle, M. Wirth, C. Papandreou, S. Bracarda, T. Marcello, C. Sternberg, C. Bangma, T. de Reijke, J. Arija, J. Bellmunt, R. Lopez, M. Lopez-Brea, A. Bjartell, J. Damber, M. Haggman, M. Hellstrom, M. Seke, J. Brown, S. Chowdhury, T. Elliott, S. Harland, H. Innes, N. James, R. Jones, D. Mazhar, E. Paez, A. Protheroe, J. Staffurth, F. Ahmann, G. Andriole, E. Arrowsmith, V. Assikis, A. Baron, W. Berry, G. Bubley, J. Carney, L. Chu, T. Cosgriff, S. Denmeade, H. Deshpande, D. Duchene, A. Ferrari, E. Frenkel, N. Gabrail, J. Garcia, D. George, L. Gomella, O. Goodman, I. Gore, J. Gullo, J. Hainsworth, O. Hamid, T. Hutson, D. King, H. Koh, A. Koletsky, F. Kudrik, P. Lara, R. Lyons, J. Maranchie, M. Modiano, J. Nieva, L. Nordquist, J. Pinski, B. Poiesz, J. Polikoff, D. Quinn, C. Redfern, S. Riggs, C. Ryan, M. Saleh, A. Sartor, M. Scholz, N. Shore, S. Srinivas, U. Vaishampaya, J. Vieweg, M. Vira, N. Vogelzang, G. Wilding, Y. Wong, A. Belldegrun, P. W. Kantoff, M. A. Carducci, N. J. Vogelzang, W. K. Kelly, R. J. Auchus, M. Meyers, W. Rackoff, N. Tran, M. Yu, R. Knoblauch, V. Naini, S. Matheny, S. Maul, J. Larsen, J. Martin, H. Wawda, D. Goffredo, J. Li, S. Li, B. Li, K. Durve, M. J. Morris, and S. M. Larson. Abiraterone in metastatic prostate cancer without previous chemotherapy. N Engl J Med, 368(2):138–148, Jan 2013.

[36] Kelvin KW Chan, Helen Guo, Sierra Cheng, Jaclyn M Beca, Ruby Redmond-Misner, Wanrudee Isaranuwatchai, Lucy Qiao, Craig Earle, Scott R Berry, James J Biagi, et al. Real-world outcomes of folfirinox vs gemcitabine and nab-paclitaxel in advanced pancreatic cancer: a population-based propensity score-weighted analysis. Cancer medicine, 9(1):160–169, 2020.

[37] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Bryan Lau, Adaora A Adimora, and Women's Interagency HIV Study. Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS research and human retroviruses, 30(12):1170–1177, 2014.

[38] Di Shu, Kazuki Yoshida, Bruce H Fireman, and Sengwee Toh. Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. Statistical methods in medical research, 29(6):1668–1681, 2020.

[39] Chongliang Luo, Rui Duan, Adam C Naj, Henry R Kranzler, Jiang Bian, and Yong Chen. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. Scientific reports, 12(1):6627, 2022.

[40] Dongdong Li, Wenbin Lu, Di Shu, Sengwee Toh, and Rui Wang. Distributed Cox proportional hazards regression using summary-level information. Biostatistics, 2022.

[41] Ji A Park, Tae H Kim, Jihoon Kim, and Yu R Park. WICOX: Weight-based integrated Cox model for time-to-event data in distributed databases without data-sharing. IEEE Journal of Biomedical and Health Informatics, 27(1):526–537, 2022.

[42] Rhian Daniel, Jingjing Zhang, and Daniel Farewell. Making apples from oranges: Comparing non-collapsible effect estimators and their standard errors after adjustment for different covariate sets. Biometrical Journal, 63(3):528–557, 2021.

[43] C. J. Ryan, M. R. Smith, K. Fizazi, F. Saad, P. F. Mulders, C. N. Sternberg, K. Miller, C. J. Logothetis, N. D. Shore, E. J. Small, J. Carles, T. W. Flaig, M. E. Taplin, C. S. Higano, P. de Souza, J. S. de Bono, T. W. Griffin, P. De Porre, M. K. Yu, Y. C. Park, J. Li, T. Kheoh, V. Naini, A. Molina, D. E. Rathkopf, A. Boyce, A. Costello, I. Davis, V. Ganju, L. Horvath, R. Lynch, P. Mainwaring, G. Marx, S. Ng, F. Parnis, J. Shapiro, N. Singhal, M. Slancar, G. Van Hazel, S. Wong, S. D. Yip, P. Carpentier, D. Luyten, S. Rottey, D. Schrijvers, F. Van Aelst, H. Van Poppel, T. Cheng, J. Chin, S. Ellard, Y. Fradet, M. Gleave, A. Joshua, L. Klotz, H. Martins, S. Mukherjee, S. North, E. Winquist, S. Abdel-Hamid, M. Colombel, A. chon, O. Haillot, F. Joly, S. Oudard, F. Priou, E. Raymond, P. Albers, M. Boegemann, J. Gleissner, J. Gschwend, P. Hammerer, A. Heidenreich, M. Kuczyk, K. Miller, R. Oetzel, J. Roigas, T. Steuber, M. ckle, H. Suttman, M. Wirth, E. Efstathiou, C. Papandreou, S. Bracarda, T. Marcello, C. Sternberg, C. Bangma, W. Gerritsen, J. Arranz Arija, J. Bellmunt, R. Lopez, M. Lopez-Brea, J. Piulats, A. Bjartell, J. Damber, M. Haggman, M. Hellstrom, M. Seke, J. Brown, S. Chowdhury, J. de Bono, T. Elliott, S. Harland, H. Innes, N. James, R. Jones, D. Mazhar, E. Paez, A. Protheroe, J. Staffurth, F. Ahmann, G. Andriole, E. Arrowsmith, V. Assikis, A. Baron, E. Basch, W. Berry, G. Bubley, J. Carney, L. Chu, T. Cosgriff, S. Denmeade, H. Deshpande, D. Duchene, E. Estathiou, A. Ferrari, E. Frenkel, N. Gabrail, J. Garcia, D. George, L. Gomella, O. Goodman, I. Gore, J. Gullo, J. Hainsworth, O. Hamid, T. Hutson, D. King, H. Koh, A. Koletsky, F. Kudrik, P. Lara, R. Lyons, J. Maranchie, M. Modiano, J. Nieva, L. Nordquist, A. Pantuck, J. Pinski, B. Polesz, J. Polikoff, D. Quinn, C. Redfern, S. Riggs, C. Ryan, M. Saleh, A. Sartor, H. Scher, M. Scholz, N. Shore, S. Srinivas, U. Vaishampaya, J. Vieweg, M. Vira, N. Vogelzang, G. Wilding, Y. Wong, and E. Yu. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. Lancet Oncol, 16(2):152–160, Feb 2015.

[44] Sara Pusceddu, Michele Ghidini, Martina Torchio, Francesca Corti, Gianluca Tomasello, Monica Niger, Natalie Prinzi, Federico Nichetti, Andrea Coinu, Maria Di Bartolomeo, et al. Comparative effectiveness of gemcitabine plus nab-paclitaxel and folfirinox in the first-line setting of metastatic pancreatic cancer: a systematic review and meta-analysis. Cancers, 11(4):484, 2019.

[45] Elena Gabriela Chiorean, Winson Y Cheung, Guido Giordano, George Kim, and Salah-Eddin Al-Batran. Real-world comparative effectiveness of nab-paclitaxel plus gemcitabine versus folfirinox in advanced pancreatic cancer: a systematic review. Therapeutic advances in medical oncology, 11:1758835919850367, 2019.

[46] Nicolas Williet, Angélique Saint, Anne-Laure Pointet, David Tougeron, Simon Pernot, Astrid Pozet, Dominique Bechade, Isabelle Trouilloud, Nelson Lourenco, Vincent Hautefeuille, et al. Folfirinox versus gemcitabine/nab-paclitaxel as first-line therapy in patients with metastatic pancreatic cancer: a comparative propensity score study. Therapeutic advances in gastroenterology, 12:1756284819878660, 2019.

[47] Avital Klein-Brill, Shlomit Amar-Farkash, Gabriella Lawrence, Eric A Collisson, and Dvir Aran. Comparison of folfirinox vs gemcitabine plus nab-paclitaxel as first-line chemotherapy for metastatic pancreatic ductal adenocarcinoma. JAMA network open, 5(6):e2216199–e2216199, 2022.

[48] S Hegewisch-Becker, A Aldaoud, T Wolf, B Krammer-Steiner, H Linde, R Scheiner-Sparna, D Hamm, M Jänicke, and N Marschner. Tpk-group (tumour registry pancreatic cancer). results from the prospective german tpk clinical cohort study: Treatment algorithms and survival of 1,174 patients with locally advanced, inoperable, or metastatic pancreatic ductal adenocarcinoma. Int J Cancer, 144(5):981–990, 2019.

[49] Jakob M Riedl, Florian Posch, Lena Horvath, Antonia Gantschnigg, Felix Renneberg, Esther Schwarzenbacher, Florian Moik, Dominik A Barth, Christopher H Rossmann, Michael Stotz, et al. Gemcitabine/nab-paclitaxel versus folfirinox for palliative first-line treatment of advanced pancreatic cancer: A propensity score analysis. European Journal of Cancer, 151:3–13, 2021.

[50] Jung Won Chun, Sang Hyub Lee, Joo Seong Kim, Namyoung Park, Gunn Huh, In Rae Cho, Woo Hyun Paik, Ji Kon Ryu, and Yong-Tae Kim. Comparison between folfirinox and gemcitabine plus nab-paclitaxel including sequential treatment for metastatic pancreatic cancer: a propensity score matching approach. BMC cancer, 21(1):537, 2021.

[51] The Lancet Gastroenterology Hepatology. Cause for concern: the rising incidence of early-onset pancreatic cancer, 2023.

[52] Lihui Zhao, Brian Claggett, Lu Tian, Hajime Uno, Marc A Pfeffer, Scott D Solomon, Lorenzo Trippa, and LJ Wei. On the restricted mean survival time curve in survival analysis. Biometrics, 72(1):215–221, 2016.

[53] Kyongsun Pak, Hajime Uno, Dae Hyun Kim, Lu Tian, Robert C Kane, Masahiro Takeuchi, Haoda Fu, Brian Claggett, and Lee-Jen Wei. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA oncology, 3(12):1692–1696, 2017.

[54] Sarah C Conner, Lisa M Sullivan, Emelia J Benjamin, Michael P LaValley, Sandro Galea, and Ludovic Trinquart. Adjusted restricted mean survival times in observational studies. Statistics in medicine, 38(20):3832–3860, 2019.

[55] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. The Journal of Machine Learning Research, 21(1):8747–8752, 2020.

[56] Irva Hertz-Picciotto and Beverly Rockhill. Validity and efficiency of approximation methods for tied survival times in Cox regression. Biometrics, pages 1151–1156, 1997.

[57] Cameron Davidson-Pilon. lifelines: survival analysis in python. Journal of Open Source Software, 4(40):1317, 2019.

[58] Bradley Efron. The efficiency of Cox's likelihood function for censored data. Journal of the American statistical Association, 72(359):557–565, 1977.

[59] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. ACM Trans. Knowl. Discov. Data, 15(5), may 2021.

[60] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. American Economic Review, 93(2):126–132, 2003.

[61] Adam Pantanowitz and Tshilidzi Marwala. Missing data imputation through the use of the random forest algorithm. In Advances in computational intelligence, pages 53–62. Springer, 2009.

[62] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

[63] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. Machine Learning, 107(8-10):1477–1494, 2018.

[64] Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. Advances in Neural Information Processing Systems, 33:5980–5990, 2020.

[65] Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. The Annals of Applied Statistics, 14(3):1409–1431, 2020.

[66] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? Advances in Neural Information Processing Systems, 34:11530–11540, 2021.

[67] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. GigaScience, 11:giac013, 2022.

[68] Na Liu, Yanhong Zhou, and J Jack Lee. IPDfromKM: reconstruct individual patient data from published kaplan-meier survival curves. BMC medical research methodology, 21(1):111, 2021.

[69] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pages 1175–1191, 2017.

[70] Burton H Bloom. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 13(7):422–426, 1970.

[71] Tanguy Marchand, Boris Muzellec, Constance Béguier, Jean Ogier du Terrail, and Mathieu Andreux. Securefedyj: a safe feature gaussianization protocol for federated learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 36585–36598. Curran Associates, Inc., 2022.

[72] Kawther Abdilleh, Omar Khalid, Dennis Ladnier, Wenshuai Wan, Sara Seepo, Garrett Rupp, Valentin Corelj, Zelia F Worman, Divya Sain, Jack DiGiovanna, et al. Pancreatic cancer action network's spark: A cloud-based patient health data and analytics platform for pancreatic cancer. JCO Clinical Cancer Informatics, 8:e2300119, 2024.

6 Online Methods

6.1 Problem statement

We consider a setting where one center, *e.g.*, a pharmaceutical company, hosts data of all treated patients and approaches several other centers to use their data as control to define a distributed ECA. Although FedECA also works in the more general case where there is no constraint on patient mixing within the participating centers.

We suppose that FDA guidelines for ECA [73] have been applied to direct data harmonization so that variables, assigned or received treatments, data formats, variable ranges, outcome definitions and inclusion

criteria match across centers. However we touch on the practical challenges associated with such a requirement when studying the metastatic pancreatic adenocarcinoma use-case in Section 6.4.3. We assume that variables are not missing and relegate discussing data imputation questions associated with real-world data to Section 6.4.3 as well.

We consider as well that all centers arrived at a consensus on a common list of confounding factors that influence both the exposure and the outcome of interest, we give examples of such lists in Section 6.4.2 and Section 6.4.3.

Moreover, the studied treatment effect is the average treatment effect (ATE) evaluated using the hazard ratio (HR) with time-to-event outcomes. See the discussion regarding this design choice.

Finally we assume the deployment of a federated solution such as Substra [74] between the centers as well as a trusted third party or aggregator. This setup is detailed for the "real-world" deployment use-cases in Section 6.5.

We note that because of the scope of this article, we do not necessarily dwell on such technicalities; in practice however, they are a crucial aspect of FL projects and should not be underestimated [75, 76, 77].

6.2 Federated External Control Arms (FedECA)

6.2.1 Method overview

The ECA methodology we use relies on 3 main steps: training a propensity score model, fitting a weighted Cox model, and testing the parameter related to the treatment. We first introduce them here in a pooled-level fashion, before explaining in detail how we adapted them to the federated setting in the next sections.

Setup and notations

Each patient is represented by covariates $\boldsymbol{X} \in \mathbb{R}^p$. It undergoes treatment $A \in \{0, 1\}$, corresponding either to the treated ($A = 1$) or control ($A = 0$) arm. We denote $\boldsymbol{x}_i$ the covariates of the i-th patient, and a_i its treatment allocation. Following treatment, the patient has an event of interest (*e.g.*, death or disease relapse) at a random time T^*. The patient may leave the arm before the event of interest is actually observed, a phenomenon called censoring: we denote the *observed* time T, whose realizations are denoted t_i. We note $\delta_i = 1$ if this corresponds to a true event, resp. $\delta_i = 0$ if censorship took place. Additionally, we define the observed outcome $Y_i = (T_i, \delta_i)$. Let n denote the total number of patients, indexed by i.

Let $\mathcal{S}$ denote the finite set of all observed times, i.e. $\mathcal{S} = \{t_i\}_{i=1}^n$. At a given time s, let $\mathcal{D}_s$ denote the set of patients with an event at this time, i.e.

$$\forall s \in \mathcal{S}, \mathcal{D}_s = \{i | t_i = s, \delta_i = 1\}, \tag{1}$$

and let $\mathcal{R}_s$ denote the set of patients *at risk* at this time, i.e.

$$\forall s \in \mathcal{S}, \mathcal{R}_s = \{i | t_i \geq s\}. \tag{2}$$

Further, let $\mathring{\mathcal{S}}$ denote the set of times where at least one true event occurs, i.e.

$$\mathring{\mathcal{S}} = \{s \in \mathcal{S} | \mathcal{D}_s \neq \emptyset\}. \tag{3}$$

Data is distributed among K different centers, with n_k samples per center. We denote $\boldsymbol{x}_{i,k}$ the i-th covariate vector from the k-th center; accordingly, $a_{i,k}$ denotes the treatment allocation, $y_{i,k} = (t_{i,k}, \delta_{i,k})$ the observed outcome, where $t_{i,k}$ is the observed time event, and $\delta_{i,k}$ whether a true event took place. Similarly, for each time s and center k, we define the subset $\mathcal{D}_{s,k}$ and $\mathcal{R}_{s,k}$ as the respective restrictions of $\mathcal{D}_s$ and $\mathcal{R}_s$ to center k.

Propensity score model training

Due to the lack of randomization, for each sample, the probability of being assigned the treatment A might depend on the covariates $\boldsymbol{X}$. We train a propensity score model $p_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ such that

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx \mathbb{P}[A | \boldsymbol{X} = \boldsymbol{x}]. \tag{4}$$

We use a logistic model for $p_{\boldsymbol{\theta}}$, i.e.,

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}. \tag{5}$$

Its negative log-likelihood is given by

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ a_i \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + (1-a_i)\log(1-p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \right\}. \tag{6}$$

In Section 6.2.3, we explain how this model is trained in a federated setting.

Inverse Probability Weighted Treatment (IPTW)

For each sample i, we define an IPTW weight $w_i \in (0, +\infty)$ based on the propensity score model trained in the previous step as

$$w_i = \begin{cases} \frac{1}{\max(p_\theta(\boldsymbol{x}_i),\varepsilon)} & \text{if } a_i = 1, \\ \frac{1}{\max(1-p_\theta(\boldsymbol{x}_i),\varepsilon)} & \text{otherwise.} \end{cases} \tag{7}$$

In order to avoid overflow errors, $\varepsilon > 0$ was set to 10^{-16} in our experiments.

We then train a weighted Cox proportional hazards (CoxPH) model with parameters $\boldsymbol{\beta} \in \mathbb{R}^q$, related to patient-specific variables $\boldsymbol{z}_i \in \mathbb{R}^q$. We stress that the variables $\boldsymbol{z}_i$ are not the same as the covariates $\boldsymbol{x}_i$. More precisely, for the vanilla IPTW method, the sole covariate used is the treatment allocation, i.e., $\boldsymbol{z}_i = a_i$. In the general case of the *adjusted* IPTW (adjIPTW) method, one may use additional covariates, especially if they are known confounders. We note that our federated framework can support both classical IPTW and adjIPTW unlike in [78], although we choose to illustrate our results with IPTW for the sake of simplicity.

The CoxPH model is fitted by maximizing a data-fidelity term consisting in the partial likelihood $L(\boldsymbol{\beta})$ with Breslow approximation [79]:

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \left(\frac{e^{\boldsymbol{\beta}^T\boldsymbol{z}_j}}{\sum\limits_{j:t_j\geq t_i} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j}} \right)^{w_i} = \prod_{s\in\mathring{\mathcal{S}}} \prod_{i\in\mathcal{D}_s} \left(\frac{e^{\boldsymbol{\beta}^T\boldsymbol{z}_i}}{\sum\limits_{j\in\mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j}} \right)^{w_i}, \tag{8}$$

where the second equation has been rewritten using the sets $\mathcal{D}_s$ and $\mathcal{R}_s$. For numerical stability, we use the negative log-likelihood $\ell(\beta) = \log L(\beta)$, which reads

$$\ell(\boldsymbol{\beta}) = -\sum_{s\in\mathring{\mathcal{S}}} \sum_{i\in\mathcal{D}_s} \left\{ w_i \boldsymbol{\beta}^T\boldsymbol{z}_i - w_i \log\left(\sum_{j\in\mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j} \right) \right\}. \tag{9}$$

While $\ell(\boldsymbol{\beta})$ represents a data-fidelity term, we also add a regularization $\psi(\boldsymbol{\beta})$ with strength $\gamma > 0$, leading to the full loss

$$\mathcal{L}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \gamma\psi(\boldsymbol{\beta}). \tag{10}$$

In Section 6.2.4, we describe how we minimize the loss $\mathcal{L}$ in a federated setting, which is the main technical innovation of this paper.

Variance estimation and statistical testing

Once the weights are fitted, we estimate the variance matrix of $\hat{\boldsymbol{\beta}}$ using a robust sandwich-type estimator [80]. Let us denote

$$\zeta_s^0(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j}, \tag{11}$$

$$\boldsymbol{\zeta}_s^1(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j} \boldsymbol{z}_j, \tag{12}$$

$$\boldsymbol{\zeta}_s^2(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T\boldsymbol{z}_j} \boldsymbol{z}_j\boldsymbol{z}_j^T, \tag{13}$$

and $\hat{\zeta}_s^0(\boldsymbol{\beta})$, $\hat{\boldsymbol{\zeta}}_s^1(\boldsymbol{\beta})$, $\hat{\boldsymbol{\zeta}}_s^2(\boldsymbol{\beta})$ the analogous quantities using the estimated weights $\{\hat{w}_i\}_{i=1}^n$.

Following [80, 78], the robust sandwich-type estimator of the variance of $\hat{\boldsymbol{\beta}}$ takes the following form:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = H^{-1}Q(H^{-1})^T, \tag{14}$$

where

$$H = \sum_{s\in\mathring{\mathcal{S}}}\sum_{i\in\mathcal{D}_s}\hat{w}_i\left(\frac{\hat{\boldsymbol{\zeta}}_s^2(\hat{\boldsymbol{\beta}})}{\hat{\zeta}_s^0(\hat{\boldsymbol{\beta}})} - \frac{\hat{\boldsymbol{\zeta}}_s^1(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\zeta}}_s^1(\hat{\boldsymbol{\beta}})^T}{\hat{\zeta}_s^0(\hat{\boldsymbol{\beta}})^2}\right), \tag{15}$$

$$Q = \sum_{i=1}^n \hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}})^T, \tag{16}$$

$$\begin{aligned}\hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}}) = \delta_i\hat{w}_i\left(\boldsymbol{z}_i - \frac{\hat{\boldsymbol{\zeta}}_s^1(\hat{\boldsymbol{\beta}})}{\hat{\zeta}_s^0(\hat{\boldsymbol{\beta}})}\right) - \hat{w}_i\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_i)\boldsymbol{z}_i\sum_{s'\in\mathring{\mathcal{S}}}\sum_{j\in\mathcal{D}_{s'}}\frac{\hat{w}_j\mathbb{1}_{\{s'\le s\}}}{\hat{\zeta}_{s'}^0(\hat{\boldsymbol{\beta}})}\\ + \hat{w}_i\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_i)\sum_{s'\in\mathring{\mathcal{S}}}\sum_{j\in\mathcal{D}_{s'}}\frac{\hat{w}_j\mathbb{1}_{\{s'\le s\}}\hat{\boldsymbol{\zeta}}_{s'}^1(\hat{\boldsymbol{\beta}})}{\hat{\zeta}_{s'}^0(\hat{\boldsymbol{\beta}})^2},\ \text{for all } i\in\mathcal{D}_s, s\in\mathring{\mathcal{S}},\end{aligned} \tag{17}$$

with $\mathbb{1}_{\{s'\le s\}}$ the indicator function that has the value 1 on all times s' (with events) and is 0 otherwise.

Eventually, a Wald test is performed on the entry of $\hat{\boldsymbol{\beta}}$ corresponding to the treatment allocation, assuming a χ^2 distribution with 1 degree of freedom [81].

6.2.2 Related works

Before diving into the details of the federation of the propensity score model and the weighted Cox model, we provide some context explaining the position of FedECA in the literature. In the case of binary or continuous outcomes, inverse probability of treatment weighting (IPTW) can be directly federated, and has been explored extensively [82, 83, 84, 85]. In contrast, to the best of our knowledge, few works have explored the federation of ML model training compatible with ECA for time-to-event outcomes.

The difficulty of this federation is that the straightforward application of FL algorithms such as Federated Averaging [86] to time-to-event ML models is impossible due to the non-separability of the Cox proportional hazards (PH) loss [87, 88]. Careful federation of the training of ML models capable of handling time-to-event outcomes is possible [87, 88] but often requires either to use tree-based models [89, 90], approximations [88, 91] or can only be performed in stratified settings [79, 78, 92, 93, 94], which limits the applicability of such federated analyses for ECA analyses. Indeed, existing *stratified* federated IPTW methods such as [78] cannot be applied to ECA as, in the realistic setting we consider, the treatment variable is constant within each center and thus comparison between the treated and untreated groups cannot be done locally from within a single center. A recent work proposed a propensity score method to estimate hazard ratios in a federated weighted Cox PH model [95]. The main difference with our work lies in the fact that they have considered propensity scores based on the combination of local propensity scores (computed in each center) and global ones demonstrating superior performance than the global scores alone.

However, as previously stated, in this paper we consider a setting where local propensity score models cannot be trained locally to predict treatment allocation since the variable to predict is constant in each center.

Furthermore, our work supports adjusted IPTW and proposes a federated algorithm for robust distributed estimation.

Other lines of work tackle the federated analytics setting where no learning is involved and propose to use aggregated data (AD), such as matching-adjusted indirect comparison (MAIC) [96] to perform direct comparisons in combination with the available individual patients data. Finally, another popular research direction is to propose private representations of patient covariates [97, 98, 99, 100] that can be pooled into a central server. These methods have the drawback of not yielding pooled-equivalent results. Furthermore, centralizing these representations increases the potential leakage risks associated with a successful, even if unlikely, attack, compared to a federated storage system. We summarize in Supplementary Table S2 the differences between the methods mentioned above.

In particular, we extend WebDISCO [87], which, alongside [95], is to the best of our knowledge, one of the few exact methods enabling federated learning of time-to-event models in ECA contexts. This extension integrates both propensity scoring and causal inference into the learning process. We also derive a federated version of the robust sandwich estimator for testing treatment effects.

Our method can also be seen as an extension of stratified IPTW for time-to-event outcomes from [78, 95] to the non-stratified case that allows application to ECA.

With this in mind, we go on to precisely describe the federation scheme by specifying all quantities that are communicated between the centers and the aggregator.

6.2.3 Federated propensity model training

Our goal is to fit a model for the propensity score (5) based on distributed data $\{(\boldsymbol{x}_{i,k}, a_{i,k})_i\}_{k=1}^K$. Let $\mathcal{J}$ denote the full negative log-likelihood of the model, and $\mathcal{J}_k$ the negative log-likehood for each center, i.e.,

$$\mathcal{J}_k(\boldsymbol{\theta}) = \sum_{i=1}^{n_k} \{a_{i,k} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{i,k}) + (1 - a_{i,k}) \log(1 - p_{\boldsymbol{\theta}}(\boldsymbol{x}_{i,k}))\} . \tag{18}$$

Due to the separability of each loss term in per-sample terms [101], we have

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathcal{J}_k(\boldsymbol{\theta}). \tag{19}$$

Using the separability (19), it is straightforward to optimize $\mathcal{J}$ using a second-order method, since its gradient and Hessian can be computed from the sum of local quantities, as described in Section 2.1 of [102]. We call this naïve strategy FedNewtonRaphson: its pseudocode is provided in Algorithm 1. This algorithm has a hyperparameter corresponding to the number of steps: in our numerical experiments, we noted that $E = 10$ is sufficient to obtain proper convergence.

The strategy FedNewtonRaphson requires to compute full batch gradients and Hessians, in time $O(n_k)$ on each center, and each communication with the aggregator requires the exchange of $O(p^2)$ floating numbers. In the setting of ECAs, we usually have both $n_k \leq 10^3$ and $p \leq 10^3$, making such a second-order approach tractable. We note that for larger data settings, several improvements could be considered following [103, 102], which would reduce the quantities of transmitted parameters. We leave such improvements to future work.

Algorithm 1 FedNewtonRaphson

Initialize $\boldsymbol{\theta}_0 = 0$
for $e = 1$ **to** E **do**
  Aggregator sends $\boldsymbol{\theta}_{e-1}$ to each center
  for $k = 1$ **to** K **in parallel do** ▷ On each center
    $\boldsymbol{g}_{e,k} = \nabla_{\boldsymbol{\theta}} J_k(\theta_{e-1})$
    $\boldsymbol{H}_{e,k} = \nabla^2_{\boldsymbol{\theta}} J_k(\theta_{e-1})$
    Send $\boldsymbol{g}_{e,k}$ and $\boldsymbol{H}_{e,k}$ to the aggregator
  end for
  $\boldsymbol{g}_e = \frac{1}{K} \sum_{k=1}^K \boldsymbol{g}_{e,k}$ ▷ Aggregator-side
  $\boldsymbol{H}_e = \frac{1}{K} \sum_{k=1}^K \boldsymbol{H}_{e,k}$
  $\boldsymbol{\theta}_e = \boldsymbol{\theta}_{e-1} - (\boldsymbol{H}_e)^{-1} \boldsymbol{g}_e$
end for
return $\boldsymbol{\theta}_E$

6.2.4 Inverse probability weighted WebDISCO

Here we propose a method to minimize the regularized weighted CoxPH model (10) in a federated fashion. Since the non-separability of the weighted CoxPH log-likelihood $\ell(\boldsymbol{\beta})$ prevents the use of vanilla FL algorithms, we inspire ourselves from WebDISCO [87] to build a pooled-equivalent second-order method.

Non-separability

Compared to the logistic propensity score model, the main difficulty of federating Equation (9) stems from the non-separability of the log-likelihood, i.e., the cross-center terms. Indeed, for any time s, the risk set $\mathcal{R}_s$ is a union of per-center terms, i.e.

$$\mathcal{R}_s = \cup_{k=1}^{K} \mathcal{R}_{s,k}. \tag{20}$$

Thus, the aggregated Equation (9) can be rewritten as

$$\ell(\boldsymbol{\beta}) = -\sum_{k=1}^{K} \sum_{s \in \mathring{\mathcal{S}}} \sum_{i \in \mathcal{D}_{s,k}} \left\{ w_i \boldsymbol{\beta}^T \boldsymbol{z}_{i,k} - w_i \log \left(\sum_{j \in \mathcal{R}_{s,k}} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_{j,k}} + \sum_{k' \neq k} \sum_{j \in \mathcal{R}_{s,k'}} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_{j,k'}} \right) \right\}, \tag{21}$$

where the loss for each sample i of each center k involves terms from other samples j in other centers $k' \neq k$. The non-separability of the CoxPH loss is a well-known issue in a federated setting and previous works have investigated reformulations to make it amenable to vanilla federated learning solvers [88]. Here we instead adapt the WebDISCO method [87] to the weighted case in order to keep pooled-equivalent results and benefit from second-order acceleration.

Federated computation of $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ and $\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$

Our method consists in performing an iterative server-level Newton-Raphson descent on $\mathcal{L}$. The gradient $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ and Hessian $\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ thus need to be computed in a federated fashion. These quantities can be computed in closed-form as

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = -\sum_{s \in \mathring{\mathcal{S}}} \sum_{i \in \mathcal{D}_s} \left(w_i \boldsymbol{z}_i - w_i \frac{\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j} \boldsymbol{z}_j}{\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j}} \right), \tag{22}$$

and

$$\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{s \in \mathring{\mathcal{S}}} \sum_{i \in \mathcal{D}_s} w_i \left\{ \frac{\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j} \boldsymbol{z}_j \boldsymbol{z}_j^T}{\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j}} - \frac{\left(\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j} \boldsymbol{z}_j \right) \left(\sum_{j' \in \mathcal{R}_s} w_{j'} e^{\boldsymbol{\beta}^T \boldsymbol{z}_{j'}} \boldsymbol{z}_{j'}^T \right)}{(\sum_{j \in \mathcal{R}_s} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j})^2} \right\}. \tag{23}$$

Note that the Hessian evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}})$, corresponds, up to a sign, to the quantity H defined in (15) for the robust variance estimator. We now define the local counterparts $\zeta^h_{s,k}(\boldsymbol{\beta})$ of the previously introduced quantities where the sum is restricted to the risk set $\mathcal{R}_{s,k}$,

$$\zeta^0_{s,k}(\boldsymbol{\beta}) = \sum_{j \in \mathcal{R}_{s,k}} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j}, \tag{24}$$

$$\boldsymbol{\zeta}^1_{s,k}(\boldsymbol{\beta}) = \sum_{j \in \mathcal{R}_{s,k}} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j} \boldsymbol{z}_j, \tag{25}$$

$$\boldsymbol{\zeta}^2_{s,k}(\boldsymbol{\beta}) = \sum_{j \in \mathcal{R}_{s,k}} w_j e^{\boldsymbol{\beta}^T \boldsymbol{z}_j} \boldsymbol{z}_j \boldsymbol{z}_j^T. \tag{26}$$

Further, let us denote

$$W_s = \sum_{i \in \mathcal{D}_s} w_i, \tag{27}$$

$$\boldsymbol{Z}_s = \sum_{i \in \mathcal{D}_s} w_i \boldsymbol{z}_i, \tag{28}$$

and

$$W_{s,k} = \sum_{i \in \mathcal{D}_{s,k}} w_i, \tag{29}$$

$$\boldsymbol{Z}_{s,k} = \sum_{i \in \mathcal{D}_{s,k}} w_i \boldsymbol{z}_i, \tag{30}$$

where by convention, in all cases, the sum is set to 0 in case of an empty set. Equations (22) and (23) can be respectively rewritten as

$$\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = -\sum_{s\in\mathring{\mathcal{S}}} \boldsymbol{Z}_s - W_s \frac{\boldsymbol{\zeta}_s^1(\boldsymbol{\beta})}{\zeta_s^0(\boldsymbol{\beta})}, \tag{31}$$

$$\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \sum_{s\in\mathring{\mathcal{S}}} W_s \left\{ \frac{\boldsymbol{\zeta}_s^2(\boldsymbol{\beta})}{\zeta_s^0(\boldsymbol{\beta})} - \frac{\boldsymbol{\zeta}_s^1(\boldsymbol{\beta})\boldsymbol{\zeta}_s^1(\boldsymbol{\beta})^T}{\zeta_s^0(\boldsymbol{\beta})^2} \right\}. \tag{32}$$

Using these equations, we can rewrite

$$\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = -\sum_{k=1}^{K} \left\{ \sum_{s\in\mathring{\mathcal{S}}} \boldsymbol{Z}_{s,k} - W_{s,k} \frac{\sum_{k'} \boldsymbol{\zeta}_{s,k'}^1(\boldsymbol{\beta})}{\sum_{k'} \zeta_{s,k'}^0(\boldsymbol{\beta})} \right\}, \tag{33}$$

$$\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{s\in\mathring{\mathcal{S}}} W_{s,k} \left\{ \frac{\sum_{k'} \boldsymbol{\zeta}_{s,k'}^2(\boldsymbol{\beta})}{\sum_{k'} \zeta_{s,k'}^0(\boldsymbol{\beta})} - \frac{\left(\sum_{k'} \boldsymbol{\zeta}_{s,k'}^1(\boldsymbol{\beta})\right)\left(\sum_{k'} \boldsymbol{\zeta}_{s,k'}^1(\boldsymbol{\beta})\right)^T}{\left(\sum_{k'} \zeta_{s,k'}^0(\boldsymbol{\beta})\right)^2} \right\}. \tag{34}$$

Assuming the set of all true event times $\mathring{\mathcal{S}}$ is known to all centers, we see that it is possible to reconstruct the full gradient $\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$ and Hessian $\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$ based on the 5-uplet $\{(W_{s,k}, \boldsymbol{Z}_{s,k}, \zeta_{s,k}^0(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^2(\boldsymbol{\beta}))\}_{s,k}$. Algorithm 2 sums up this algorithm.

Algorithm 2 FedCoxComp

Require: Weights $\boldsymbol{\beta}$, set $\mathring{\mathcal{S}}$
Aggregator sends $\boldsymbol{\beta}$ to each center
for $k = 1$ **to** K **in parallel do** ▷ On each center
  for $s \in \mathring{\mathcal{S}}$ **do**
    Compute $W_{k,s}$ with (29) ▷ 0 if $\mathcal{D}_{s,k} = \emptyset$
    Compute $\boldsymbol{Z}_{k,s}$ with (30).
  end for
  for $s \in \mathring{\mathcal{S}}$ s.t. $W_{k,s} > 0$ **do** ▷ 0 otherwise
    Compute $\zeta_{s,k}^0(\boldsymbol{\beta})$ with (24)
    Compute $\boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta})$ with (25)
    Compute $\boldsymbol{\zeta}_{s,k}^2(\boldsymbol{\beta})$ with (26)
  end for
  Send back $\{(W_k, \boldsymbol{Z}_k, \zeta_{s,k}^0(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^2(\boldsymbol{\beta}))\}_{s\in\mathring{\mathcal{S}}}$
end for
Compute $\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$ with (33) ▷ On the server
Compute $\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$ with (34)
return $\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}), \nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$

Algorithm 3 non-robust FedECA

Require: Maximal number of steps E, LR schedule $(\alpha_e)_e$, regularization γ
Initialization $\boldsymbol{\beta}_0 = 0$
for $e = 1$ **to** E **do**
  $\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_{e-1}), \nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_{e-1}) = \text{FedCoxComp}(\boldsymbol{\beta}_{e-1})$ ▷ Communication between server and centers
  $\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{e-1}) = \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_{e-1}) + \gamma\nabla_{\boldsymbol{\beta}}\psi(\boldsymbol{\beta})$
  $\nabla^2_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{e-1}) = \nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_{e-1}) + \gamma\nabla^2_{\boldsymbol{\beta}}\psi(\boldsymbol{\beta})$
  $\boldsymbol{\beta}_e = \boldsymbol{\beta}_{e-1} - \alpha_e \left(\nabla^2_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{e-1})\right)^{-1} \nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{e-1})$
  if Stopping criterion **then** $e = E$
  end if
end for
return $\boldsymbol{\beta}_E$

Non-robust FedECA

To optimize the full loss (10), we can now leverage the computation of the gradient and Hessian of the weighted CoxPH loss ℓ to perform a second-order Newton-Raphson descent. We follow the hyperparameters of `lifelines` [104] for this optimization. In particular, we use the same learning rate strategy, the same regularizer and the same stopping criterion. Indeed, as `lifeline`'s regularizer does not depend on data and is smooth, its gradient and Hessian can be computed on the server's side deriving twice the following equation:

$$\mathcal{L}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \gamma\psi(\boldsymbol{\beta}), \tag{35}$$

$$\tag{36}$$

with γ the strength of the regularization.

In more details for the regularizer $\psi(\boldsymbol{\beta})$, we use a soft elastic-net regularization [105] with hyperparameters $\lambda > 0$ and $\alpha > 0$:

$$\psi(\boldsymbol{\beta}) = \lambda \left(\sum_r \phi_\alpha(\boldsymbol{\beta}_r) \right) + \frac{1-\lambda}{2} \|\boldsymbol{\beta}\|_2^2, \tag{37}$$

where ϕ_α is a smooth approximation of the absolute value that is progressively sharpened with the round e.

$$\alpha = 1.3^e, \tag{38}$$

$$\phi_\alpha(x) = \frac{1}{\alpha} \left(\log(1 + \exp(\alpha x)) + \log(1 + \exp(-\alpha x)) \right). \tag{39}$$

We also allow for constant learning rates as in `scikit-survival` [106]. We note that implementing different learning rate strategies or regularizers should be straightforward with our implementation.

Algorithm 3 summarizes the full algorithm used.

6.2.5 Statistical test and robust variance estimation

The robust sandwich-type estimator can be obtained by aggregating local quantities as we demonstrate in the following. We assume that each client has access to $\zeta_s^0(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\zeta_s^1}(\hat{\boldsymbol{\beta}})$ for all $s \in \mathring{\mathcal{S}}$. This can be achieved by simply allowing the server to transmit the quantities $\zeta_{s,k}^0(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\zeta_{s,k}^1}(\hat{\boldsymbol{\beta}})$ to the centers in addition to H.

The global goal is to compute the robust estimator of the variance given by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = H^{-1} Q (H^{-1})^T, \tag{40}$$

where H (15) corresponds to the Hessian $\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}})$ and Q is defined in (16). We note that through FedECA (3) each client already has access to H.

Let us define M_k as

$$M_k = \sum_{i=1}^{n_k} (H^{-1} \hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}})) \hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}})^T (H^{-1})^T \ , \tag{41}$$

where the sum is on all indices belonging to client k.

Then we have,

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \sum_{k=1}^{K} M_k. \tag{42}$$

Moreover, let $\Phi(\hat{\beta}) \in \mathbb{R}^{n,p}$ be the matrix whose rows are the $\boldsymbol{\varphi}_i(\hat{\boldsymbol{\beta}})$ for all $i \in [\![1, n]\!]$. Thus we can write the variance as

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = H^{-1} \Phi(\hat{\beta})^T \Phi(\hat{\beta}) (H^{-1})^T, \tag{43}$$

$$\Phi(\hat{\beta})^T \Phi(\hat{\beta})_{i,j} = \sum_{k=1}^{n} \left(\boldsymbol{\varphi}_k(\hat{\boldsymbol{\beta}}) \right)_i \cdot \left(\boldsymbol{\varphi}_k(\hat{\boldsymbol{\beta}}) \right)_j = \sum_{k=1}^{K} \sum_{m=1}^{n_k} \left(\boldsymbol{\varphi}_m(\hat{\boldsymbol{\beta}}) \right)_i \cdot \left(\boldsymbol{\varphi}_m(\hat{\boldsymbol{\beta}}) \right)_j, \tag{44}$$

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = H^{-1} \Phi(\hat{\boldsymbol{\beta}})^T \Phi(\hat{\boldsymbol{\beta}}) (H^{-1})^T = \sum_{k=1}^{K} M_k. \tag{45}$$

Each client can compute $\boldsymbol{\varphi}_i(\hat{\boldsymbol{\beta}})$ with Eq. (17) for all its samples i ($\forall s, i \in \mathcal{D}_{s,k}$) as long as it has access to $\zeta_{s,k}^0(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\zeta}_{\boldsymbol{s,k}}^{\boldsymbol{1}}(\hat{\boldsymbol{\beta}})$ for all $s \in \mathring{\mathcal{S}}$. Therefore each client can compute the corresponding M_k.

This leads us to the full robust algorithm of FedECA in 5. Once the variance is estimated using the above expression, we can perform inference using, *e.g.*, a Z-test. Note that as in `lifelines` [104] we use the Hessian of the regularized function. Therefore to accomodate the computation of the variance we modify non-robust FedECA as depicted in Alg. 5. Privacy-wise this modification (a) gives each client the same knowledge as the server on the last round and (b) communicates an additional M_k matrix by center, which is reasonable. In addition, in the IPTW case the matrix only the treatment allocation is used as a covariate and hence M_k is a scalar.

Algorithm 4 RobustFedCoxComp

Require: Weights $\boldsymbol{\beta}$, set $\mathring{\mathcal{S}}$

Aggregator sends $\boldsymbol{\beta}$ to each center
for $k = 1$ **to** K **in parallel do** ▷ On each center
  for $s \in \mathring{\mathcal{S}}$ **do**
    Compute $W_{k,s}$ with (29) ▷ 0 if $\mathcal{D}_{s,k} = \emptyset$
    Compute $\boldsymbol{Z}_{k,s}$ with (30).
  end for
  for $s \in \mathring{\mathcal{S}}$ s.t. $W_{k,s} > 0$ **do** ▷ 0 otherwise
    Compute $\zeta_{s,k}^0(\boldsymbol{\beta})$ with (24)
    Compute $\boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta})$ with (25)
    Compute $\boldsymbol{\zeta}_{s,k}^2(\boldsymbol{\beta})$ with (26)
  end for
  Send back $\{(W_k, \boldsymbol{Z}_k, \zeta_{s,k}^0(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^2(\boldsymbol{\beta}))\}_{s \in \mathring{\mathcal{S}}}$
end for
Compute $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ with (33) ▷ On the server
Compute $\nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta})$ with (34)
return $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}), \nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta})$ ▷ **And if it's the last round return** $\forall s \in \mathring{\mathcal{S}}, \zeta_{s,k}^0(\boldsymbol{\beta}), \boldsymbol{\zeta}_{s,k}^1(\boldsymbol{\beta}), W_s$

Algorithm 5 FedECA

Require: Maximal number of steps E, LR schedule $(\alpha_e)_e$, regularization γ

Initialization $\boldsymbol{\beta}_0 = 0$
for $e = 1$ **to** E **do**
  $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}_{e-1}), \nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta}_{e-1}) = \text{RobustFedCoxComp}(\beta_{e-1})$ ▷ Communication between server and centers
  $\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{e-1}) = \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}_{e-1}) + \gamma \nabla_{\boldsymbol{\beta}} \psi(\boldsymbol{\beta})$
  $\nabla_{\boldsymbol{\beta}}^2 \mathcal{L}(\boldsymbol{\beta}_{e-1}) = \nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta}_{e-1}) + \gamma \nabla_{\boldsymbol{\beta}}^2 \psi(\boldsymbol{\beta})$
  $\boldsymbol{\beta}_e = \boldsymbol{\beta}_{e-1} - \alpha_e \left(\nabla_{\boldsymbol{\beta}}^2 \mathcal{L}(\boldsymbol{\beta}_{e-1}) \right)^{-1} \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{e-1})$
  if Stopping criterion **then** $e = E$
  end if
end for
return $\boldsymbol{\beta}_E$
Define $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_E$
for $k = 1$ **to** K **in parallel do** ▷ On each center
  Send back $M_k M_k^T$ where $M_k = \sum_{s \in \mathring{\mathcal{S}}} \sum_{i \in \mathcal{D}_{s,k}} H^{-1} \hat{\boldsymbol{\varphi}}_i(\hat{\boldsymbol{\beta}})$.
end for
Compute $\widehat{Var}(\hat{\boldsymbol{\beta}}) = \sum_{k=1}^K M_k M_k^T$ ▷ On the server
return $\widehat{Var}(\hat{\boldsymbol{\beta}})$

6.2.6 Federated bootstrap

When implementing bootstrap in federated setups, the naïve way is to bootstrap samples per-center. However this creates some edge cases if centers have small sample sizes (*e.g.* $n_k = 1$) and risks underestimating the variance compared to the pooled case. We therefore implement both but use in our experiments a global bootstrap where we sample with replacement the global distributed cohort as if it were pooled.

6.2.7 Privacy of FedECA

We consider that time-to-event and censorship are safe to share, this is a strong assumption but is often used in clinical trials as KM curves are released [107].

Regarding the security of the covariates, we place ourselves in the *honest-but-curious* threat model, described in more detail in Substra's documentation [108].

The only covariate used when doing IPTW is the treatment allocation, which is known throughout centers. Therefore the only quantities tied to the covariates that are communicated are (1) the gradients of the propensity model, and (2) the scalar product of covariates and propensity model weights that are exposed through the propensity scores, averaged on risk sets and on distinct event times. Regarding the first point we propose an implementation of a differentially private version of the propensity model training that we describe in the next paragraph. Regarding the second point we assume that the dimension p of the covariate vector is such that $p >> 1$ and therefore that leaking scalar products is an acceptable risk in this context; This is a strong assumption. In the general case it could theoretically allow for attacks such as membership attacks [109]. Making the pipeline end-to-end differential private (DP) is an open problem. One could in principle rely again on DP to either add noise to the scalar products themselves or to the propensity scores when training the Cox PH model. However, this would affect the result even more than when applying DP only to the propensity model training, which already has a strong effect see Supplementary Figure S2. Another research avenue would be to increase the average/minimum size of the per-client risk sets by discretizing the times and applying random quantization mechanisms (RQM) [110]. We note that in this second case another downside would be that, in addition to destabilizing the training of the Cox model, it would artificially create more ties in the data, which would in return affect the quality of the Breslow estimator.

Because it exposes sums of additional covariates, studying the privacy of the federation of *adjusted* IPTW requires a specific treatment that we leave to future work.

Differential privacy of the propensity model

We list here some properties of DP that are relevant to our implementation and refer the reader to the work of [111] or [112] for a more complete exposition of the topic:

DP provides *slack* parameters (ϵ, δ), which allow to strike a trade-off between model accuracy and privacy of individual contributions.

A process $\mathcal{M}$ is (ϵ, δ)-DP if and only if $\forall D, D'$ adjacent (differing by one element), we have:

$$p(\mathcal{M}(D) \in S) \leq p(\mathcal{M}(D') \in S) \cdot \exp(\epsilon) + \delta \tag{46}$$

Perfect privacy guarantees are only obtained by taking $(\epsilon, \delta) = (0, 0)$ which makes the process $\mathcal{M}$ provably indistinguishable from the addition or removal of one individual. In practice in real-world deployments it seems ϵ between 0.1 and 50 are used depending on the application [112] with different values of δ. DP benefits from nice composability properties [111] and can thus be applied easily to ML training methods that are iterative by nature and can therefore be applied to FL as well [113].

We use the Opacus library [114], which implements the privacy accountant method of [113] to train the propensity model within FedECA with differential privacy (DP) with various (ϵ, δ) couples.

Our implementation is available in the script torch_dp_fed_avg_algo.py and uses Rényi differential privacy (RDP) [115] which gives tighter bounds alongside with Poisson sampling.

6.3 Federated Analytics

Regulators ask for SMD and Kaplan-Meier survival curves [116] in addition to the hazard ratio (HR), the associated confidence intervals (CI) and p-value in order to validate the reweighting of the propensity model and to be able to describe the patient population's time-to-events distribution within the two groups.

Therefore we also implement two federated analytics methods that we call Fed-Kaplan and Fed-SMD in order to compute such quantities globally on distributed data without compromising data.

It is to be noted that Fed-Kaplan can be directly derived from FedECA as it requires the same quantities, namely the risk sets and number of occurence of events. However, Fed-SMD requires the communication of additional second order terms which increases the attack surface of FedECA.

Those two implementations are novel to the best of our knowledge.

6.3.1 Federated Kaplan-Meier estimator

We follow FedECA implementation to compute per-center and communicate the unique times of events $\mathcal{S}_k$, the (weighted) risk set $\mathcal{R}_{s,k}$ and the (weighted) number of deaths occurring at these times $\mathcal{D}_{s,k}$ for each arm.

This enables to compute in the server $\mathcal{R}_s$ and $\mathcal{D}_s$ which then allows to compute the Kaplan-Meier estimator at each time t of a predefined grid for each arm as well as the Greenwood and exponential Greenwood confidence intervals [117].

For completeness, we remind the reader of this well-known formulas that we rewrite using our notations:

$$\hat{S}(t) = \prod_{s \in \mathring{\mathcal{S}} | s \leq t} \left(1 - \frac{\sum_{j \in \mathcal{R}_s} w_j}{\sum_{k \in \mathcal{D}_s} w_k}\right),$$

$$\mathrm{Var}(\hat{S}) = \hat{S}(t)^2 \prod_{s \in \mathring{\mathcal{S}} | s \leq t} \frac{\sum_{j \in \mathcal{D}_s} w_j}{(\sum_{k \in \mathcal{R}_s} w_k) \times \left(\sum_{k \in \mathcal{R}_s} w_k - \sum_{j \in \mathcal{D}_s} w_j\right)},$$

$$Z(t) = \log(\log \hat{S}(t)),$$

$$\mathrm{Var}[\hat{Z}(t)] = \frac{1}{(\log \hat{S}(t))^2} \prod_{s \in \mathring{\mathcal{S}} | s \leq t} \frac{\sum_{j \in \mathcal{D}_s} w_j}{(\sum_{k \in \mathcal{R}_s} w_k) \times \left(\sum_{k \in \mathcal{R}_s} w_k - \sum_{j \in \mathcal{D}_s} w_j\right)}.$$

With $\hat{S}(t)$ the Kaplan-Meier estimator of the survival function and $\mathrm{Var}(\hat{S})$ and $\mathrm{Var}[\hat{Z}(t)]$ respectively the Greenwood and exponential Greenwood estimators of the variance of the Kaplan-Meier estimator at time t. In practice exponential Greenwood shall be used [117] and this is what we display in the results.

6.3.2 SMD estimator

6.3.2.1 Computing SMD in a federated setting

We compute the standardized mean difference (SMD) for each covariate before and after weighting as defined by:

$$SMD = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}. \tag{47}$$

Where $\bar{x}_1$ and $\bar{x}_2$ are the means of the covariate in the two arms and s_1^2 and s_2^2 are the variances of the covariate in the two arms. As explained in [118, 119] we use the variance of the groups before weighting as a normalizer. We compute this quantity in a federated fashion efficiently in two aggregation rounds by developing the variance following [120]:

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \tag{48}$$

Effectively each center transmits uncentered moments of order 1 and 2 and the server uses them to derive the centered moments of order 1 and 2.

6.4 Datasets and cohorts construction

6.4.1 Synthetic data generating model of time-to-event outcome

To illustrate the performance of our proposed FL implementation, we rely on simulations with synthetic data. We simulate covariates and related time-to-event outcomes respecting the proportional hazards (PH) assumption, with the baseline hazard function derived from a Weibull distribution. For simplicity we assume a constant treatment effect across the population. The data generation process consists of several consecutive steps that we describe below assuming our target is a dataset with p covariates and n samples.

First, a design matrix $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(p)}] \in \mathbb{R}^{n \times p} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ is drawn from a multivariate normal distribution to obtain (baseline) observations for n individuals described by p covariates. The covariance matrix $\boldsymbol{\Sigma}$ is taken to be a Toeplitz matrix such that the covariances between pairs $(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)})$ of covariates decay geometrically. In other words, for a fixed $\rho > 0$, we have $\text{cov}(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}) = \rho^{|i-j|}$. Such a covariance matrix implies a locally and hierarchically grouped structure underlying the covariates, which we choose to mimic the potentially complex structure of real-world data. To reflect the varying correlations of the covariates with the outcome of interest, the coefficients $\boldsymbol{\beta}_i$ of the linear combination used to build the hazard ratio are drawn from a standard normal distribution.

$$\begin{aligned} &\boldsymbol{\Sigma} = \text{Toeplitz}(1, \rho, \rho^2, \cdots, \rho^{p-1}), \\ &\boldsymbol{X} \in \mathbb{R}^{n \times p} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \\ &\boldsymbol{\beta} \in \mathbb{R}^{p} \sim \mathcal{N}(0, 1). \end{aligned} \tag{49}$$

In the context of clinical trials with external control arms, which implies non-randomized treatment allocation, we simulate the treatment allocation in such a way that it depends on the covariates. More precisely, we introduce the treatment allocation variable A that follows a Bernoulli distribution, where the probability of being treated (the propensity score) q depends on a linear combination of the covariates, connected by a logit link function g. The coefficients $\boldsymbol{\alpha}_i$ of the linear combination are drawn from a uniform distribution, where the range $k \geq 0$ is symmetric around 0 and is normalized by the number of covariates. The degree of influence of the covariates on A can be regulated by adjusting the value of k. The greater the value of k, the stronger the influence, and therefore the lower the degree of overlap between the distributions of propensity scores of the treated and (external) control groups. Conversely, $k = 0$ removes the dependence, leading to a randomized treatment allocation.

$$\begin{aligned} &\boldsymbol{\alpha} \in \mathbb{R}^{p} \sim p^{-1/2} U(-k, k), \\ &q_i = g^{-1}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) = (1 + e^{-\boldsymbol{\alpha}^T \boldsymbol{X}_i})^{-1}, \\ &a_i | \boldsymbol{X}_i \sim \text{Bern}(q_i). \end{aligned} \tag{50}$$

Once drawn, the treatment allocation variable A_i is composed with the constant treatment effect, defined here as the hazard ratio μ, to obtain the final hazard ratio h_i for each individual. The time-to-event T_i^* of each sample is then drawn from a Weibull distribution with shape ν and the scale depending on h_i and ν. Meanwhile, for all samples we assume a constant dropout (or censoring) rate d across time, resulting in a censoring time that follows an exponential distribution.

$$\begin{aligned} &h_i(a_i) = \mu^{a_i} \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i), \\ &T_i^* \sim \mathcal{W}(h_i(a_i)^{-\frac{1}{\nu}}, \nu), \\ &C_i \sim \mathcal{E}(d) \end{aligned} \tag{51}$$

Finally, the event indication variable δ_i can be derived from T_i^* and C_i: $\delta_i = \mathbb{1}_{T_i^* \leq C_i}$. And the observed outcome Y_i for the ith individual is defined as the couple $Y_i = (T_i = \min(T_i^*, C_i), \delta_i)$, i.e., it corresponds to the observed time and the information on whether an event is observed.

6.4.2 Prostate cancer cohort construction

We access data of two phase III randomized clinical trials from the Yale University Open Data Access (YODA) project [121, 122] of patients with metastatic castration-resistant prostate cancer. The first trial, NCT02257736, has apalutamide, abiraterone acetate and prednisone (Apa-AA-P) for the treatment arm, and placebo, abiraterone acetate and prednisone (AA-P) for the control arm. The primary outcome is radiographic progression-free survival (rPFS). The second trial, NCT00887198, has AA-P for the treatment arm, and placebo and prednisone (P) for the control arm. The primary outcomes are overall survival (OS) and rPFS. We thus artificially distribute the data of the first trial to one "client" (simulated server) and the data of the second arm to the second "client" replicating natural splits such as in [91] to simulate a federated learning setup.

While all patients are randomized in each trial, it is still necessary to correct for potential confounding when comparing arms from different trials. First, the inclusion/exclusion criteria of both trials were aligned, patients in NCT02257736 with present visceral metastases at randomization were excluded to match the exclusion criterion of NCT00887198. Then a group of variables of patient's baseline characteristics were chosen for propensity-weighting based on literature review as well as on their availability in both trials. The chosen covariates are age, body-mass index (BMI), eastern cooperative oncology group (ECOG), brief pain inventory (BPI) score and bone-metastasis-only. We present baseline characteristics for each trial in Supplementary Table S3 and Supplementary Table S4.

We then filter these patients to remove non-informative patients and patients with missing survival information. The final full cohort consists of $n = 1927$ patients ($n = 839$ for NCT02257736 and $n = 1088$ for NCT00887198) in three treatment arms (Apa-AA-P, AA-P and P). We infer the missing covariates on a per-center basis using MissForest [123]. The flow diagram of the cohort construction is present in Supplementary Figure S4, including different ECA experiments conducted in this study.

We note that, since we submitted our research plan proposal to YODA (provided in Supplementary Figure S5) in order to access the data, we departed from the original plan in the following ways:

- IPTW is studied instead of G-computation
- we do not study conformal prediction
- time-to-event endpoints are studied instead of change in SLD or change in PSA

6.4.3 Pancreatic adenocarcinoma cohort construction

Cohort construction

We access data from three different sources: the Fédération Francophone de Cancérologie Digestive (FFCD), the Institut d'Investigació Biomèdica de Girona (IDIBGI), and the Pancreatic Cancer Action Network (PanCAN). The FFCD data consists of a subset of two clinical trials: PRODIGE 35 [124] and PRODIGE 37 [125] that respectively compare the first line efficacy of, for PRODIGE 35, 6 months of FOLFIRINOX (arm A), 4 months of FOLFIRINOX followed by leucovorin plus fluorouracil maintenance treatment for controlled patients (arm B), and a sequential treatment alternating gemcitabine and fluorouracil, leucovorin, and irinotecan every 2 months (arm C) and for PRODIGE 37: alternately receive gemcitabine + nab-paclitaxel for 2 months then FOLFIRI.3 for 2 months in arm A, or gemcitabine + nab-paclitaxel alone until progression in arm B. We use both the FOLFIRINOX arm B ($n = 92$) and the gemcitabine + nab paclitaxel arm A from PRODIGE 37 with ($n = 61$). The inclusion criteria of this new subset is thus metastatic pancreatic adenocarcinoma patients with a performance status eastern cooperative oncology group (ECOG) of either 0, 1 or 2. We select patients with the same inclusion criteria treated with FOLFIRINOX or gemcitabine + nab-paclitaxel from clinical practice data from IDIBGI and PanCAN. In PanCAN we find $n = 101$ patients treated with FOLFIRINOX and $n = 94$ with gemcitabine + nab-paclitaxel patients that meet the criteria totalling $n = 195$ patients out of 199 originally available excluding ECOG 3 and 4. We note that in PanCan, 2 patients are censored at the time the study starts therefore their data is not informative for the Cox model fitting but might still be useful for the estimation of the propensity model. In IDIBGI we find $n = 22$ patients treated with FOLFIRINOX and $n = 144$ with gemcitabine + nab-paclitaxel.

For each patient we access the following covariates: age at diagnosis, ECOG performance status, biological gender and whether or not patients have liver metastasis following the literature [126] and restrictions due to

data availability for each covariate in each center. We present baseline characteristics for each of the centers in Supplementary Table S7,Supplementary Table S5 and Supplementary Table S6 respectively for FFCD, IDIBGI and PanCAN.

We then filter these patients to remove non-informative patients, i.e., patients with missing treatment or survival information.

The final full distributed cohort consists of $n = 514$ patients ($n = 153$ for FFCD, $n = 166$ for IDIBGI and $n = 195$ for PanCAN).

We infer the missing covariates on a per-center basis using MissForest [123] considering ECOG as a numerical variable because it is ordered and apply minimum-maximum normalization to numerical variables using $[0, 100]$ for age values and $[0.0, 2.0]$ for ECOG loosely following [126].

We provide below the inclusion and ethics statements related to the above data.

Inclusion and ethics statement. The ethics of this retrospective study on clinical data collected during care and from past clinical trials were validated for each institution according to corresponding local regulations. We list below the corresponding statements from each of the participating cancer centers.

Regarding FFCD data, the study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) requirements and Good Clinical Practice guidelines; it received authorization from the French national medicines agency (ASNM), and independent ethics committee (number 214-R18 and 14-12-79 respectively for PRODIGE 35 and PRODIGE 37). The study was both registered in clinicaltrials.gov (NCT02352337 for PRODIGE 37 and NCT02827201 for PRODIGE 35) and EudraCT 2014-004449-28.

For IDIBGI data the study was approved by the Comitè d'Ètica d'Investigació amb Medicaments CEIM GIRONA the 8th of August 2023 (Acta 11/2023) under reference CEIM code 2023.165 with principal investigators ADELAIDA GARCIA VELASCO and ROBERT CARRERAS TORRES and SANOFI-AVENTIS SA as promoter.

Finally for PanCAN data the sponsor of the IRB was the Pancreatic Cancer Action Network (# KYT001) the IRB reference is 20192301, study 1265508 with main investigator Matrisian, Lynn.

Practical considerations associated with setting-up real-world federated learning collaborations

While clinical trials data is well-standardized, real-world data from centers from multiple continents are not and need to be harmonized for the federation to be considered. Clinical practice data has to be extracted by partners from different local sources stored in different databases and accessed by different internal toolings leading to a variety of extracted formats. We ask the centers to align on a common data dictionary created from FFCD data, which acts as the reference center as RCT data is already well-curated. We share this dictionary as a Google Sheet to all partners specifying expected variables, units formats and possible values. Resulting data extracts have missing values and some data entries contain errors. Thus, while some parts can be automated, the whole process from data extraction to data harmonization involves some back and forth between data engineers from partner centers, medical doctors, data stewarts and data scientists in order to perform thorough quality checks of the input data. It is interesting to note that, while we did not use large language models (LLMs) [127] in this work, they could certainly be useful to streamline parts of this process [128], However, end-to-end automation seems out of reach with current technology [129].

6.5 Real-world experiments setup details

All experiments in this article are simulated in-RAM with the exception of two experiments: the first one which uses synthetic data and splits it into 10 cloud nodes and the pancreatic adenocarcinoma use-case.

We refer to those two experiments as "real-world" in order to distinguish the complexity of their deployment from in-RAM simulation cases.

In both cases, we use the Substra platform [74] to deploy the federated learning network over secure cloud-based infrastructures.

Substra is distributed with Helm charts for each component. The charts package all the files required for a deployment in a Kubernetes cluster. Provisioning of the clusters and Substra deployment are performed using a private Terraform module (known as infrastructure as-code). We detail below the two different deployments.

6.5.1 Synthetic data infrastructure setup

For this experiment, the clusters are hosted on Google Kubernetes engine (GKE) but Substra's deployment is cloud-agnostic. Provisioning of the GKE cluster and Substra deployment are performed using a private Terraform module (known as infrastructure as-code). For this experiment, we used 11 Kubernetes clusters:

- 1 cluster is hosting the Substra orchestrator - single source of truth within the federation - as well as a Substra Backend and Frontend, which makes it capable of receiving and performing aggregation tasks. Substra's documentation refers to this cluster as "AggregationNode".
- 10 clusters are hosting a Substra Backend (and Frontend) only ; performing compute tasks on local data. Substra's documentation refers to each of these clusters as "TrainDataNode".

Clusters are physically in Belgium according to Google ("zone europe-west1" [*]). GKE version used is `1.27.2-gke.1200` and the machines used are the "n1-standard-16" [†]. Regarding the communication protocol between centers, the organizations communicate with the orchestrator via gRPC and over http(s) one to another. Since the experiment is simulated in an internal environment using synthetic data we chose not to enforce mutual transport layer security (mTLS). More information can be found in Substra's documentation[‡].

6.5.2 Metastatic pancreatic adenocarcinoma data infrastructure setup

Similarly we deploy within Owkin Inc.'s own Federated Research Network cloud infrastructure four nodes: a node for each of the three participating centers and an additional server node the "AggregationNode" with responsibilities described above. Each node here is independent: the nodes are deployed in different regions with different cloud providers. PanCan data are located in the US while Idibgi and FFCD data are hosted in Europe. The precise version information of the Substra versions used for this deployment are `0.51.0+dev` for `substra-frontend`, `0.47.0+d5dfbdb6` for `substra-backend` and `0.42.0+e6b1bddb` for the `orchestrator` repository.

Partner centers uploaded their data to the corresponding nodes.

6.6 Estimation of the treatment effect

We compare FedECA to several competitors. The unweighted Cox regression and the MAIC methods are suited for the distributed data setting if we assume that the times and events of patients can be shared across centers, while the Pooled IPTW method is only applicable on pooled data. Given the time-to-event nature of the outcome, we choose to estimate the hazard ratio under the proportional hazards assumption as a measure of the treatment effect. For all competitors, data is used to fit a Cox model as implemented in the `lifelines` library [104] to obtain the estimation.

Unweighted Cox regression

We implement a naïve Cox model regressing the observed outcome Y on the treatment allocation variable A, without using the weights of the samples. This corresponds to an unadjusted comparison between the treated and untreated groups, which would be valid in a randomized setting but not in an external control arm case. This estimator corresponds to the WebDISCO method and we use the implementation provided by the authors of this method.

MAIC

To compare FedECA to the MAIC method, we make use of the implementation available in the `indcomp` package[§]. The two methods differ mainly in the way in which samples are reweighted. More specifically, for the MAIC method, individual patient data samples (IPD) are reweighted so that a specified group of covariates matches the external aggregated data in terms of means and variances, whereas the samples in the

[*] `https://cloud.google.com/compute/docs/regions-zones?hl=en`

[†] `https://cloud.google.com/compute/docs/general-purpose-machines?hl=en#n1_machine_types`

[‡] `https://docs.substra.org/en/latest/documentation/components.html`

[§] `https://github.com/AidanCooper/indcomp`

external aggregated data are unweighted (uniform weight of 1). This creates by design reweighted data with an SMD of 0 for each covariate. The two data sources are then combined to fit a Cox model incorporating the observed outcome Y and the treatment allocation A, taking into account the reweighted results.

Pooled IPTW

The general concept and strategy of IPTW has been described before (see Section 6.2.1). In the implementation, the core estimation process is divided into two key steps. First, the propensity scores are estimated using unpenalized logistic regression or, alternatively, they can be provided externally to the estimator. These scores are then used to compute inverse probability weights tailored to the effect estimand. For the average treatment effect (ATE), weights are based on the inverse of propensity scores for both treated and control groups. For the average treatment effect on the treated (ATT), the weights involve a combination of treatment indicators (for the treated individuals) and inverse propensity scores (for the control individuals). Second, the treatment effect estimation is performed by fitting a weighted Cox proportional hazards model, where the inverse probability weights are incorporated in the regression model of the observed outcome Y on the treatment allocation A.

Competing paradigms

We already motivated the choice of IPTW as the best weighting method for small sample sizes. However other methods than weighting and matching could be considered for federation as well such as G-computation [130, 131] or doubly debiased machine learning [132, 133] as their performance should be comparable [133]. We leave their federation to future work.

6.7 Experiments details

6.7.1 Efficient bootstrap implementation with Substra

Distributed computation with Substra introduces an overhead per atomic task executed locally on each partner's machine. This overhead is not negligible and can be a bottleneck when performing bootstrapping, which naïvely necessitate to execute $O(n_{rounds} * n_{bootstraps})$ tasks per-client. To alleviate this issue we implement a more efficient bootstrapping strategy where we only execute $O(n_{rounds})$ tasks. This is achieved by employing hooks so that each task instead bootstraps itself and then executes each round. This requires to also modify the aggregation to be able to aggregate each bootstrap run separately. Details of this non-trivial implementation trick can be found in the bootstraper.py script.

Note that we could also add another layer of parallelization inside each task by using Python multiprocessing as each bootstrap run is independent of each other. The impact of this optimization would be negligible with respect to the overhead introduced by the distributed constraints. This way, a Substra experiments with 200 bootstraps lasts less than an hour instead of $\approx$ 200 hours naïvely without this parallelization layer; note that other kinds of parallelization schemes could also be undertaken such as running multiple training jobs (so-called "Compute Plans" in Substra) in parallel as was done in MELLODDY [134]. However this option necessitate scaling servers' computational ressources (CPUs, RAM) linearly with the number of training jobs in parallel which is impractical.

6.7.2 Early stopping within a static distributed framework

As Substra is static and requires to fix the number of federated rounds *a priori*, we implement early-stopping for the stopping criterion on the Hessian norm by running up to MAX_{iter} rounds (20 in practice) and backtrack to find the first round where convergence was achieved.

6.7.3 Software and Reproducibility

Following the recent trend of switching from R to Python for implementing statistical software [106, 135, 136], we choose Python as the base language for our implementation. This choice is also motivated by the fact that most FL research implementation code is written in Python. We follow reference survival analysis packages implementation design choices such as `lifelines` [104] and `scikit-survival` [106]. We use the Substra

software [74] which is an open-source software that has been audited and validated by security teams of both hospitals and pharmaceutical companies across different FL projects. Substra has demonstrated its ability to be deployed in real-world conditions for biomedical research purposes in the MELLODDY project [134, 134], as well as in the HealthChain project on breast cancer treatment response prediction [137].

FedECA is available as a Python package on Github[¶] for non-commercial use.

The code in this repository follows best practices such as continuous integration (CI), thorough code testing (coverage of code at 82% on commit `480fa75`), deployed documentation using Github pages as well as the use of pre-commit hooks to help manage the repository's evolution.

The availability of the code not only ensures the reproducibility of the results presented in this article as well as the possibility to audit its implementation, but also opens the possibility for other research teams to perform real-world federated ECA.

Indeed, a user can launch FedECA running the exact same code either in-RAM for simulations, or on a real deployed substra network in real conditions by modifying the backend type, as shown in Listing Supplementary Listing S1.

The FedECA repository contains a quickstart as well as detailed documentation and comments, which should allow easy replication.

All quantitative figures in this article with synthetic data can be reproduced by following instructions in `experiments/README.md`. The associated yaml configurations provide all hyper-parameters that were used.

For experiments on 10 centers replication involves deploying a substra network, which require some development operations (DevOps) capabilities. However details in section 6.5.1 should be sufficient to reproduce the results. The associated experiment script is defined in real_world_runtimes.yaml.

For experiments on YODA data, we install the fedeca package within the YODA platform, split the data in such a way that the control arm and the treatment arm are in two separate groups, and run fedeca with bootstrap variance estimation. Scripts used to preprocess the data and run the experiments are available in the yoda folder in the `fedeca` repository.

For experiments on metastatic pancreatic adenocarcinoma data, we use the fedeca package unaltered on commit `b6474e5` after having registered the data in the Substra platform. Obfuscated versions of the scripts that ran on the deployed platform and that were used to generate the related figures in the article are available in the pdac folder in the `fedeca` repository. Where by "obfuscated" we mean that dataset hashes or urls in this script were converted to random strings so they cannot be mapped to any of the original data or servers.

The nature of this last experiment is such that replications require data access which might be restricted, see Section 7. However once access to data is obtained and federated network is deployed all experiments should be easily reproduced thanks to the above scripts.

Further questions can be addressed to the corresponding author J.O.d.T. through the creation of github issues or via direct e-mail.

Methods-only references

[73] US Food, Drug Administration, et al. Considerations for the design and conduct of externally controlled trials for drug and biological products. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products`, February 2023.

[74] Mathieu N Galtier and Camille Marini. Substra: a framework for privacy-preserving, traceable and collaborative machine learning. arXiv preprint arXiv:1910.11567, 2019.

[75] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2):1–210, 2021.

[¶] `https://github.com/owkin/fedeca`

[76] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. Nature communications, 13(1):7346, 2022.

[77] Markus R Bujotzek, Ünal Akünal, Stefan Denner, Peter Neher, Maximilian Zenk, Eric Frodl, Astha Jaiswal, Moon Kim, Nicolai R Krekiehn, Manuel Nickel, et al. Real-world federated learning in radiology: Hurdles to overcome and benefits to gain. arXiv preprint arXiv:2405.09409, 2024.

[78] Di Shu, Kazuki Yoshida, Bruce H Fireman, and Sengwee Toh. Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. Statistical methods in medical research, 29(6):1668–1681, 2020.

[79] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Bryan Lau, Adaora A Adimora, and Women's Interagency HIV Study. Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS research and human retroviruses, 30(12):1170–1177, 2014.

[80] David A Binder. Fitting Cox's proportional hazards models from survey data. Biometrika, 79(1):139–147, 1992.

[81] John P Klein, Melvin L Moeschberger, et al. Survival analysis: techniques for censored and truncated data, volume 1230. Springer, 2003.

[82] Sengwee Toh, Robert Wellman, R Yates Coley, Casie Horgan, Jessica Sturtevant, Erick Moyneur, Cheri Janning, Roy Pardee, Karen J Coleman, David Arterburn, et al. Combining distributed regression and propensity scores: a doubly privacy-protecting analytic method for multicenter research. Clinical Epidemiology, pages 1773–1786, 2018.

[83] Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. arXiv preprint arXiv:2107.11732, 2021.

[84] Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. arXiv preprint arXiv:2112.09313, 2021.

[85] Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. arXiv preprint arXiv:2309.12600, 2023.

[86] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017.

[87] Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. WebDISCO: a web service for distributed Cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association, 22(6):1212–1219, 2015.

[88] Mathieu Andreux, Andre Manoel, Romuald Menuet, Charlie Saillard, and Chloé Simpson. Federated survival analysis with discrete-time Cox models. arXiv preprint arXiv:2006.08997, 2020.

[89] Xuan Wang, Harrison G Zhang, Xin Xiong, Chuan Hong, Griffin M Weber, Gabriel A Brat, Clara-Lea Bonzel, Yuan Luo, Rui Duan, Nathan P Palmer, et al. SurvMaximin: Robust federated approach to transporting survival risk prediction models. Journal of biomedical informatics, 134:104176, 2022.

[90] Alberto Archetti and Matteo Matteucci. Federated survival forests. arXiv preprint arXiv:2302.02807, 2023.

[91] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. Advances in Neural Information Processing Systems, 35:5315–5334, 2022.

[92] Chongliang Luo, Rui Duan, Adam C Naj, Henry R Kranzler, Jiang Bian, and Yong Chen. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. Scientific reports, 12(1):6627, 2022.

[93] Dongdong Li, Wenbin Lu, Di Shu, Sengwee Toh, and Rui Wang. Distributed Cox proportional hazards regression using summary-level information. Biostatistics, 2022.

[94] Ji A Park, Tae H Kim, Jihoon Kim, and Yu R Park. WICOX: Weight-based integrated Cox model for time-to-event data in distributed databases without data-sharing. IEEE Journal of Biomedical and Health Informatics, 27(1):526–537, 2022.

[95] Chen Huang, Kecheng Wei, Ce Wang, Yongfu Yu, and Guoyou Qin. Covariate balance-related propensity score weighting in estimating overall hazard ratio with distributed survival data. BMC medical research methodology, 23(1):233, 2023.

[96] James E. Signorovitch, Vanja Sikirica, M. Haim Erder, Jipan Xie, Mei Lu, Paul S. Hodgkins, Keith A. Betts, and Eric Q. Wu. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. Value in Health, 15(6):940–947, 2012.

[97] Jeremy A Rassen, Jerry Avorn, and Sebastian Schneeweiss. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. Pharmacoepidemiology and drug safety, 19(8):848–857, 2010.

[98] Junghye Lee, Jimeng Sun, Fei Wang, Shuang Wang, Chi-Hyuck Jun, Xiaoqian Jiang, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR medical informatics, 6(2):e7744, 2018.

[99] Yuji Kawamata, Ryoki Motai, Yukihiko Okada, Akira Imakura, and Tetsuya Sakurai. Collaborative causal inference on distributed data. arXiv preprint arXiv:2208.07898, 2022.

[100] Akira Imakura, Ryoya Tsunoda, Rina Kagawa, Kunihiro Yamagata, and Tetsuya Sakurai. DC-COX: Data collaboration Cox proportional hazards model for privacy-preserving survival analysis on multiple parties. Journal of Biomedical Informatics, 137:104264, 2023.

[101] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. arXiv preprint arXiv:1812.02903, 2018.

[102] Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. In International conference on machine learning, pages 4617–4628. PMLR, 2021.

[103] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Feddane: A federated newton-type method. In 2019 53rd Asilomar Conference on Signals, Systems, and Computers, pages 1227–1231. IEEE, 2019.

[104] Cameron Davidson-Pilon. lifelines: survival analysis in python. Journal of Open Source Software, 4(40):1317, 2019.

[105] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2):301–320, 2005.

[106] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. The Journal of Machine Learning Research, 21(1):8747–8752, 2020.

[107] Na Liu, Yanhong Zhou, and J Jack Lee. IPDfromKM: reconstruct individual patient data from published kaplan-meier survival curves. BMC medical research methodology, 21(1):111, 2021.

[108] Substra Foundation. Privacy strategy - substra documentation, 2023. Accessed: 2024-09-17.

[109] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pages 3–18. IEEE, 2017.

[110] Yeojoon Youn, Zihao Hu, Juba Ziani, and Jacob Abernethy. Randomized quantization is all you need for differential privacy in federated learning. arXiv preprint arXiv:2306.11913, 2023.

[111] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.

[112] Damien Desfontaines. A list of real-world uses of differential privacy. `https://desfontain.es/privacy/real-world-differential-privacy.html`, 10 2021. Ted is writing things (personal blog).

[113] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.

[114] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. arXiv preprint arXiv:2109.12298, 2021.

[115] Ilya Mironov. Rényi differential privacy. In 2017 IEEE 30th computer security foundations symposium (CSF), pages 263–275. IEEE, 2017.

[116] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282):457–481, 1958.

[117] S Sawyer. The greenwood and exponential greenwood confidence intervals in survival analysis. Applied survival analysis: regression modeling of time to event data, pages 1–14, 2003.

[118] Noah Greifer. Covariate balance tables and plots: A guide to the 'cobalt' package. `https://ngreifer.github.io/cobalt/articles/cobalt.html#variance-in-standardized-mean-differences-and-correlations`, 2023. Accessed: 2024-08-30.

[119] Peter C Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in medicine, 27(12):2037–2049, 2008.

[120] Philippe Pébay, Timothy B Terriberry, Hemanth Kolla, and Janine Bennett. Numerically stable, scalable formulas for parallel and online computation of higher-order multivariate central moments with arbitrary weights. Computational Statistics, 31:1305–1325, 2016.

[121] Harlan M Krumholz and Joanne Waldstreicher. The yale open data access (yoda) project–a mechanism for data sharing. The New England journal of medicine, 375(5):403–405, 2016.

[122] Joseph S Ross, Joanne Waldstreicher, Stephen Bamford, Jesse A Berlin, Karla Childers, Nihar R Desai, Ginger Gamble, Cary P Gross, Richard Kuntz, Richard Lehman, et al. Overview and experience of the yoda project with clinical trial data sharing after 5 years. Scientific data, 5(1):1–14, 2018.

[123] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

[124] Laetitia Dahan, Nicolas Williet, Karine Le Malicot, Jean-Marc Phelip, Jérôme Desrame, Olivier Bouché, Caroline Petorin, David Malka, Christine Rebischung, Thomas Aparicio, et al. Randomized phase ii trial evaluating two sequential treatments in first line of metastatic pancreatic cancer: results of the panoptimox-prodige 35 trial. Journal of Clinical Oncology, 39(29):3242–3250, 2021.

[125] Yves Rinaldi, Anne-Laure Pointet, Faiza Khemissa Akouz, Karine Le Malicot, Bidaut Wahiba, Samy Louafi, Alain Gratet, Laurent Miglianico, Hortense Laharie, Karine Bouhier Leporrier, et al. Gemcitabine plus nab-paclitaxel until progression or alternating with folfiri. 3, as first-line treatment for patients with metastatic pancreatic adenocarcinoma: The federation francophone de cancérologie digestive-prodige 37 randomised phase ii study (firgemax). European Journal of Cancer, 136:25–34, 2020.

[126] Avital Klein-Brill, Shlomit Amar-Farkash, Gabriella Lawrence, Eric A Collisson, and Dvir Aran. Comparison of folfirinox vs gemcitabine plus nab-paclitaxel as first-line chemotherapy for metastatic pancreatic ductal adenocarcinoma. JAMA network open, 5(6):e2216199–e2216199, 2022.

[127] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[128] Jakob Nikolas Kather, Dyke Ferber, Isabella C Wiest, Stephen Gilbert, and Daniel Truhn. Large language models could make natural language again the universal interface of healthcare. Nature Medicine, pages 1–3, 2024.

[129] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. Nature, 620(7972):172–180, 2023.

[130] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical modelling, 7(9-12):1393–1512, 1986.

[131] Arthur Chatton, Florent Le Borgne, Clémence Leyrat, Florence Gillaizeau, Chloé Rousseau, Laetitia Barbin, David Laplaud, Maxime Léger, Bruno Giraudeau, and Yohann Foucher. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Scientific reports, 10(1):9219, 2020.

[132] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[133] Nicolas Loiseau, Paul Trichelair, Maxime He, Mathieu Andreux, Mikhail Zaslavskiy, Gilles Wainrib, and Michael GB Blum. External control arm analysis: an evaluation of propensity score approaches, g-computation, and doubly debiased machine learning. BMC Medical Research Methodology, 22(1):1–13, 2022.

[134] Martijn Oldenhof, Gergely Ács, Balázs Pejó, Ansgar Schuffenhauer, Nicholas Holway, Noé Sturm, Arne Dieckmann, Oliver Fortmeier, Eric Boniface, Clément Mayer, Arnaud Gohier, Peter Schmidtke, Ritsuya Niwayama, Dieter Kopecky, Lewis Mervin, Prakash Chandra Rathi, Lukas Friedrich, András Formanek, Peter Antal, Jordon Rahaman, Adam Zalewski, Wouter Heyndrickx, Ezron Oluoch, Manuel Stößel, Michal Vančo, David Endico, Fabien Gelus, Thaïs de Boisfossé, Adrien Darbier, Ashley Nicollet, Matthieu Blottière, Maria Telenczuk, Van Tien Nguyen, Thibaud Martinez, Camille Boillet, Kelvin Moutet, Alexandre Picosson, Aurélien Gasser, Inal Djafar, Antoine Simon, Ádám Arany, Jaak Simm, Yves Moreau, Ola Engkvist, Hugo Ceulemans, Camille Marini, and Mathieu Galtier. Industry-scale orchestrated federated learning for drug discovery. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[135] Q. Bertrand, Q. Klopfenstein, P.-A. Bannier, G. Gidel, and M. Massias. Beyond l1: Faster and better sparse models with skglm. In NeurIPS, 2022.

[136] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Bioinformatics, 39(9):btad547, 2023.

[137] Jean Ogier du Terrail, Armand Leopold, Clément Joly, Constance Béguier, Mathieu Andreux, Charles Maussion, Benoît Schmauch, Eric W Tramel, Etienne Bendjebbar, Mikhail Zaslavskiy, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature medicine, 29(1):135–146, 2023.

7 Data availability

Synthetic data used in this study can be re-generated using the scripts provided.

For the first example with real patient data, data access should be requested to the Yale University Open Data Access (YODA) Project. Detailed procedure to request data access is provided on YODA website https://yoda.yale.edu/how-request-data.

For the second example with the federated research network of FFCD, IDIBGI and PanCAN, the data is under restricted access and are not freely available as specific clearance from the ethics committee and compliance with local regulations is required to access data from each center. Data from PanCAN is available to qualified researchers by submitting a proposal for review at https://spark.sbgenomics.com/. Data from FFCD and IDIBGI can also be made available upon reasonable request. Specific conditions and restrictions of access to the datasets are to be discussed directly with the main investigators in each center: J.-B. B. for FFCD, R. C. for IDIBGI.

8 Code Availability

The integrality of the code is publicly released and openly available for research purposes at the following URL: `https://github.com/owkin/fedeca` at the exception of the code used to launch metastatic pancreatic adenocarcinoma which will be released upon acceptance.

Supplementary Material

Method	Environment	#centers	Runtime (s)
FedECA (robust)	real-world setup	2	$4.48 \cdot 10^3 \pm 1.08 \cdot 10^2$
FedECA (robust)	real-world setup	3	$4.57 \cdot 10^3 \pm 5.88 \cdot 10^1$
FedECA (robust)	real-world setup	5	$4.58 \cdot 10^3 \pm 9.42 \cdot 10^1$
FedECA (robust)	real-world setup	8	$4.56 \cdot 10^3 \pm 9.53 \cdot 10^1$
FedECA (robust)	real-world setup	10	$5.00 \cdot 10^3 \pm 7.80 \cdot 10^2$
FedECA (robust)	in-RAM	2	$4.95 \pm 8.21 \cdot 10^{-1}$
FedECA (robust)	in-RAM	3	$6.72 \pm 4.92 \cdot 10^{-1}$
FedECA (robust)	in-RAM	5	$1.27 \cdot 10^1 \pm 1.72$
FedECA (robust)	in-RAM	8	$1.43 \cdot 10^1 \pm 1.38$
FedECA (robust)	in-RAM	10	$1.93 \cdot 10^1 \pm 2.02$
IPTW	–	–	$2.34 \cdot 10^{-1} \pm 2.41 \cdot 10^{-2}$

Supplementary Table S1: Runtimes of different federated and pooled experiments in different conditions: in-RAM simulations or running in a deployed Substra network in the cloud (real-world setup).

```python
from fedeca import FedECA
from fedeca.survival_utils import CoxData

SEED = 42  # Seed for the generation of synthetic data
NSAMPLES = 1000  # Number of samples in total
N_CLIENTS = 5  # Number of simulated centers
N_COV = 10  # Number of covariates
# Types of backend used for the FL, simu is the most lightweight,
# real-world FL is "remote"
BACKEND_TYPE = "simu"
# Simulates FL by splitting a dataframe across centers and register
# each dataset into Substra. In case of a real deployment, private
# datasets are registered by each organization's data engineers.
data = CoxData(seed=SEED, n_samples=NSAMPLES, ndim=N_COV)
df = data.generate_dataframe()
df.drop(columns=["propensity_scores"], axis=1, inplace=True)
# As in sklearn we first instantiate an object
fedeca = FedECA(N_COV, treated_col="treated", duration_col="T", event_col="E",
                num_rounds_list=[50, 50], variance_method="robust")
# We then call the fit method of the object to launch the FL
fedeca.fit(df, n_clients=N_CLIENTS, backend_type=BACKEND_TYPE)

```

Supplementary Listing S1: Python code to launch FedECA on simulated data using any type of deployment.

Supplementary Table S2: Comparison of distributed ECA methods for time-to-event outcomes and generic data pooling alternative. Green color highlights methods compatible with distributed ECA (ATE: average treatment effect; ATT: average treatment effect on the treated; ATC: average treatment effect on the control; KM-type information: Kaplan-Meier-type information consisting of observed time, censorship status and potentially group assignment).

	Methods								
	FedECA (Ours)	**WebDISCO** [87]	**IPW Cox** [78]	**FCI** [83]	**DC-COX** [100]	**ODACH** [92]	**WICOX** [94]	**Huang *et al.*** [95]	**MAIC** [96]
Supported training settings									
- *stratified* Cox model training **w/o weights**	✓	✓	✓	✓	✓	✓	✓	✓	✓
- **weighted** *stratified* Cox model training	✓	✗	✓	✓	✓	✗	✗	✓	✓
- *stratified* **IPTW** analysis	✓	✗	✓	✓	✓	✗	✗	✓	✓
- *distribution-independent* Cox model training **w/o weights**	✓	✓	✗	✗	✓	✗	✗	✓	✓
- **weighted** *distribution-independent* Cox model training	✓	✗	✗	✗	✓	✗	✗	✓	✓
- *distribution-independent* **IPTW** analysis	✓	✗	✓	✓	✓	✗	✗	✓	✓
- Pooled equivalence with classical **IPTW**	✓	✓	✓	✓	✗	✓	✓	✗	✗
Treatment groups									
- Treatment groups in distinct centers	✓	✓	✗	✗	✓	✗	✗	✓	✓
- Correction for confounding	✓	✗	✓	✓	✓	✗	✗	✓	✓
- Multiple external control centers	✓	✓	✗	✗	✓	✗	✗	✓	✗
Causal estimands									
- ATE	✓	✗	✓	✓	✓	✗	✗	✓	✗
- ATT	✓	✗	✓	✓	✓	✗	✗	✓	✗
- ATC	✓	✗	✓	✓	✓	✗	✗	✓	✓
Privacy									
- Does not require pooling data	✓	✓	✓	✓	✗	✓	✓	✓	✓
- Only KM-type information is shared	✓	✓	✓	✓	✗	✓	✓	✓	✓
- Differential privacy	✓	✗	✗	✗	✗	✗	✗	✗	✗
Available implementation	✓	✓	✓	✗	✗	✓	✗	✗	✓

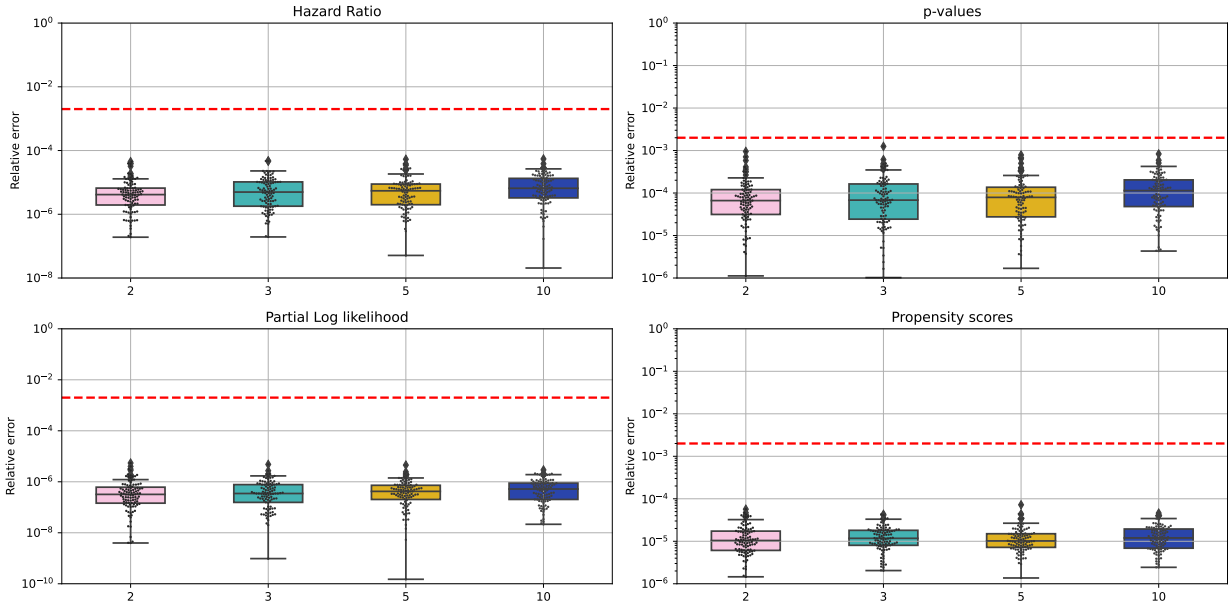
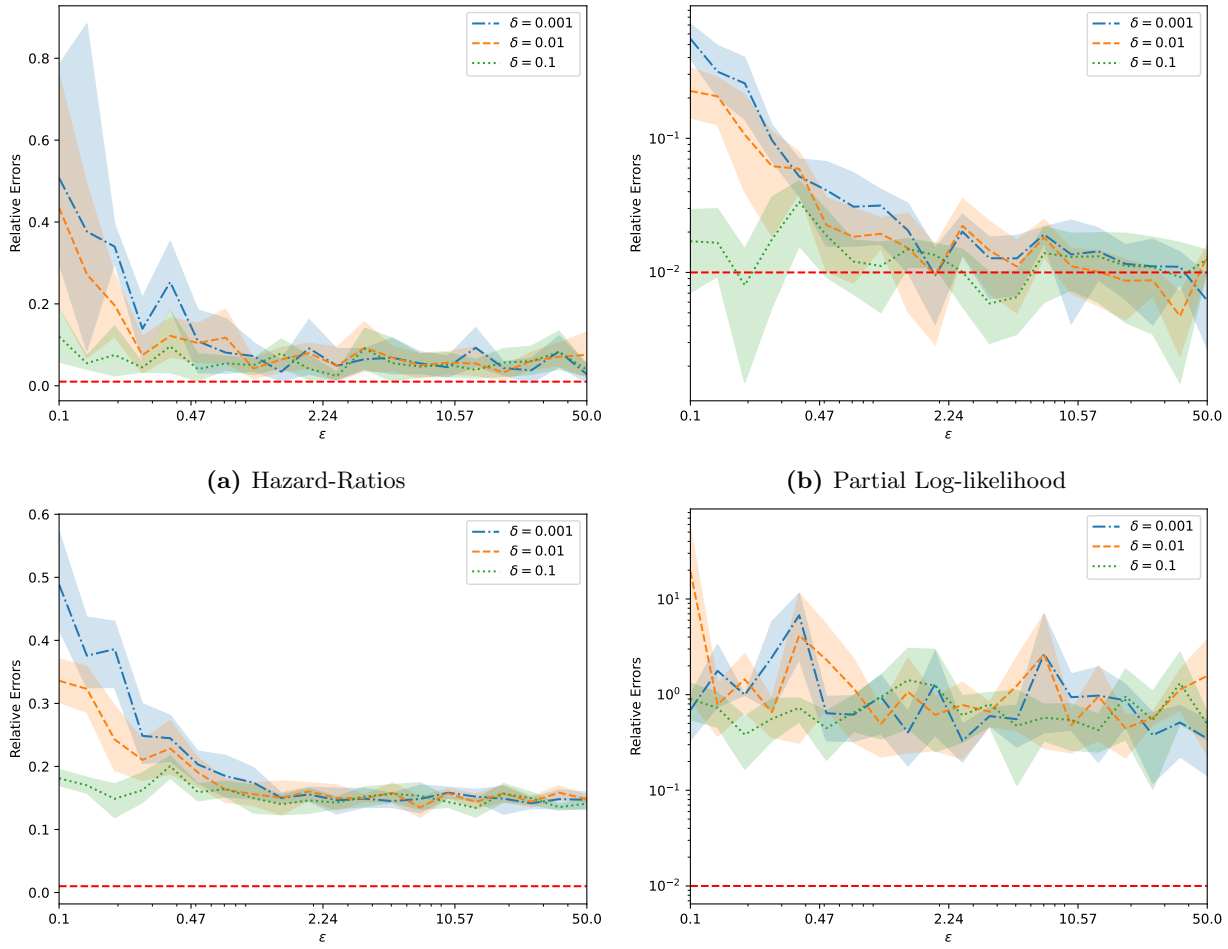


Figure Supplementary Figure S1. Pooled equivalent with varying number centers. Boxplots of the relative error between the pooled IPTW and the FedECA algorithm on four different quantities. The propensity scores estimated from the logistic regression, the hazard ratio (the treatment effect) the p-values associated to the treatment allocation variable (Wald test) and the partial likelihood resulting from the Cox model. Each of these quantities was monitored as we increased the number of centers across which the data is split from 2 to 10 centers. The errors were computed on simulated data with 100 repetitions. The red dotted line represents a relative error of 1% between pooled IPTW and FedECA.

	Apa-AA-P (n=418)	AA-P (n=421)	All (n=839)
Age at cancer diagnosis Median (IQR)	71.0 (66.0 - 78.0)	71.0 (65.0 - 77.0)	71.0 (66.0 - 77.0)
Age at cancer diagnosis Percent missing	0.0%	0.0%	0.0%
Body Mass Index (BMI) Median (IQR)	27.5 (24.9 - 30.9)	27.9 (25.1 - 31.3)	27.72 (24.9 - 31.14)
Body Mass Index (BMI) Percent missing	2.1%	2.9%	2.5%
Performance Status (ECOG) at cancer diagnosis 0	290 (69.4%)	299 (71.0%)	589 (70.2%)
Performance Status (ECOG) at cancer diagnosis 1	128 (30.6%)	122 (29.0%)	250 (29.8%)
Performance Status (ECOG) at cancer diagnosis 2	0 (0%)	0 (0%)	0 (0%)
Performance Status (ECOG) at cancer diagnosis Percent missing	0 (0%)	0 (0%)	0 (0%)
Brief Pain Inventory (BPI) score ≤ 1	317 (75.8%)	295 (70.1%)	612 (72.9%)
Brief Pain Inventory (BPI) score > 1	95 (22.7%)	118 (28.0%)	213 (25.4%)
Brief Pain Inventory (BPI) score Percent missing	6 (1.4%)	8 (1.9%)	14 (1.7%)
Bone metastasis only False	211 (50.5%)	216 (51.3%)	427 (50.9%)
Bone metastasis only True	207 (49.5%)	205 (48.7%)	412 (49.1%)
Bone metastasis only Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)

Supplementary Table S3: Baseline characteristics of NCT02257736.



(a) Hazard-Ratios

(b) Partial Log-likelihood

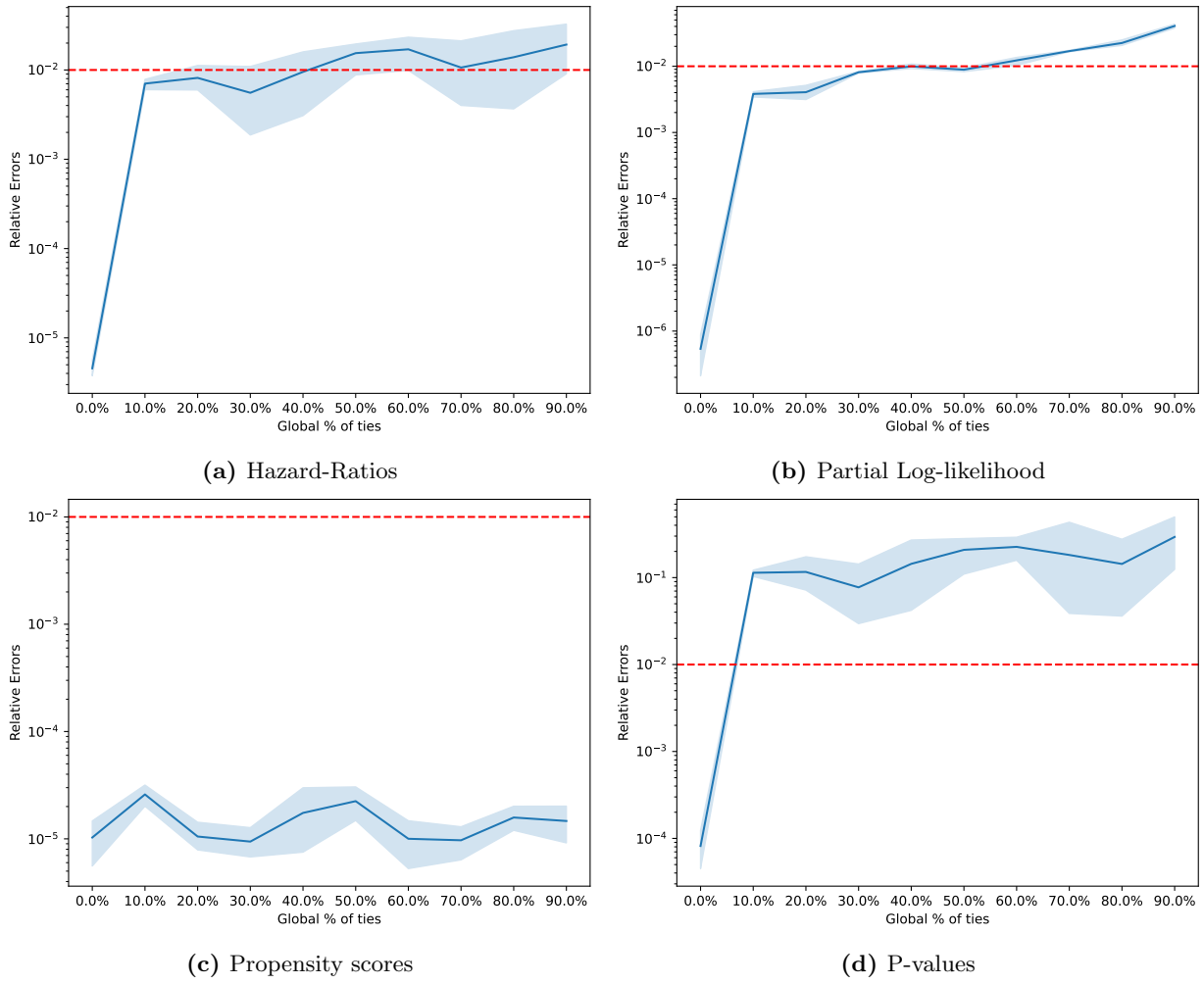
(c) Propensity scores

(d) P-values

Figure Supplementary Figure S2. DP-FedECA. Adding differential privacy into FedECA. Comparison of the results of running DP-FedECA with respect to the pooled baseline with no privacy. We see that even for large ϵ that correspond to lower amount of noise, the relative difference between the p-values produced by DP-FedECA and the true p-value is high even if the propensity weights are relatively close. The final operation to build the p-value involves a second-order term which is very sensitive to the precise value of the propensity scores.

	AA-P (n=546)	P (n=542)	All (n=1088)
Age at cancer diagnosis Median (IQR)	68.0 (64.0 - 76.0)	68.0 (60.0 - 76.0)	68.0 (64.0 - 76.0)
Age at cancer diagnosis Percent missing	0.0%	0.0%	0.0%
Body Mass Index (BMI) Median (IQR)	28.4 (26.0 - 31.5)	28.4 (25.8 - 31.8)	28.4 (25.8 - 31.6)
Body Mass Index (BMI) Percent missing	2.2%	2.2%	2.2%
Performance Status (ECOG) at cancer diagnosis 0	411 (75.3%)	409 (75.5%)	820 (75.4%)
Performance Status (ECOG) at cancer diagnosis 1	134 (24.5%)	133 (24.5%)	267 (24.5%)
Performance Status (ECOG) at cancer diagnosis 2	1 (0.2%)	0 (0%)	1 (0.1%)
Performance Status (ECOG) at cancer diagnosis Percent missing	0 (0%)	0 (0%)	0 (0%)
Brief Pain Inventory (BPI) score ≤ 1	370 (67.8%)	346 (63.8%)	716 (65.8%)
Brief Pain Inventory (BPI) score > 1	129 (23.6%)	147 (27.1%)	276 (25.4%)
Brief Pain Inventory (BPI) score Percent missing	47 (8.6%)	49 (9%)	96 (8.8%)
Bone metastasis only False	286 (52.4%)	281 (51.8%)	567 (52.1%)
Bone metastasis only True	238 (43.6%)	241 (44.5%)	479 (44.0%)
Bone metastasis only Percent missing	22 (4.0%)	20 (3.7%)	42 (3.9%)

Supplementary Table S4: Baseline characteristics of NCT00887198.



(a) Hazard-Ratios

(b) Partial Log-likelihood

(c) Propensity scores

(d) P-values

Figure Supplementary Figure S3. Influence of ties on FedECA accuracy. Comparison of the results of running FedECA with respect to the pooled baseline using Efron's approximation. Performance degrades with the number of ties. For realistic number of ties errors are manageable <1%.

	FOLFIRINOX (n=22)	Gemcitabine + Nab-Paclitaxel (n=144)	All (n=166)
Age at cancer diagnosis	57.27 (37.00 - 75.00)	65.10 (37.00 - 86.00)	64.07 (37.00 - 86.00)
Age at cancer diagnosis Percent missing	0.0%	0.0%	0.0%
Performance Status (ECOG) at cancer diagnosis 0	12 (54.5%)	32 (22.2%)	44 (26.5%)
Performance Status (ECOG) at cancer diagnosis 1	8 (36.4%)	85 (59.0%)	93 (56.0%)
Performance Status (ECOG) at cancer diagnosis 2	1 (4.5%)	15 (10.4%)	16 (9.6%)
Performance Status (ECOG) at cancer diagnosis Percent missing	1 (4.5%)	12 (8.3%)	13 (7.8%)
Biological gender M	14 (63.6%)	76 (52.8%)	90 (54.2%)
Biological gender F	8 (36.4%)	68 (47.2%)	76 (45.8%)
Biological gender Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)
Liver metastasis True	17 (77.3%)	104 (72.2%)	121 (72.9%)
Liver metastasis False	5 (22.7%)	40 (27.8%)	45 (27.1%)
Liver metastasis Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)

Supplementary Table S5: Baseline characteristics of IDIBGI.

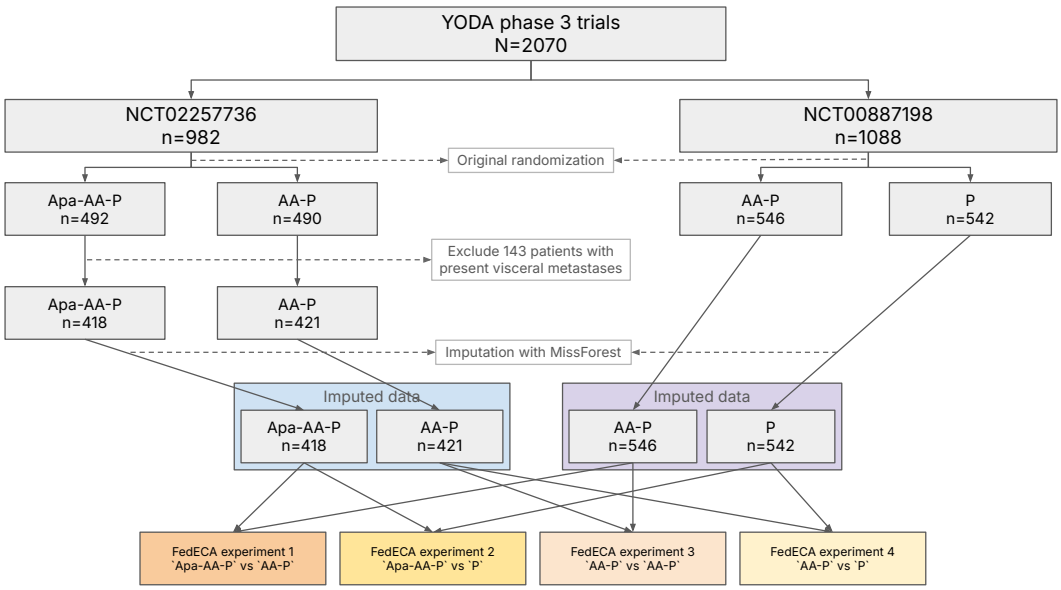


Figure Supplementary Figure S4. Flow diagram of YODA clinical trial cohort construction.

	FOLFIRINOX (n=101)	Gemcitabine + Nab-Paclitaxel (n=94)	All (n=195)
Age at cancer diagnosis	61.41 (38.00 - 78.00)	63.86 (40.00 - 84.00)	62.60 (38.00 - 84.00)
Age at cancer diagnosis Percent missing	1.0%	0.0%	0.5%
Performance Status (ECOG) at cancer diagnosis 0	30 (29.7%)	21 (22.3%)	51 (26.2%)
Performance Status (ECOG) at cancer diagnosis 1	43 (42.6%)	48 (51.1%)	91 (46.7%)
Performance Status (ECOG) at cancer diagnosis 2	3 (3.0%)	8 (8.5%)	11 (5.6%)
Performance Status (ECOG) at cancer diagnosis Percent missing	25 (24.8%)	17 (18.1%)	42 (21.5%)
Biological gender M	69 (68.3%)	51 (54.3%)	120 (61.5%)
Biological gender F	32 (31.7%)	43 (45.7%)	75 (38.5%)
Biological gender Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)
Liver metastasis False	23 (22.8%)	30 (31.9%)	53 (27.2%)
Liver metastasis True	78 (77.2%)	64 (68.1%)	142 (72.8%)
Liver metastasis Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)

Supplementary Table S6: Baseline characteristics of PanCAN.

	FOLFIRINOX (n=92)	Gemcitabine + Nab-Paclitaxel (n=61)	All (n=153)
Age at cancer diagnosis	62.64 (39.90 - 75.97)	64.12 (40.97 - 75.95)	63.23 (39.90 - 75.97)
Age at cancer diagnosis Percent missing	0.0%	0.0%	0.0%
Performance Status (ECOG) at cancer diagnosis 0	36 (39.1%)	23 (37.7%)	59 (38.6%)
Performance Status (ECOG) at cancer diagnosis 1	56 (60.9%)	31 (50.8%)	87 (56.9%)
Performance Status (ECOG) at cancer diagnosis 2	0 (0.0%)	7 (11.5%)	7 (4.6%)
Performance Status (ECOG) at cancer diagnosis Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)
Biological gender F	36 (39.1%)	33 (54.1%)	69 (45.1%)
Biological gender M	56 (60.9%)	28 (45.9%)	84 (54.9%)
Biological gender Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)
Liver metastasis False	16 (17.4%)	10 (16.4%)	26 (17.0%)
Liver metastasis True	76 (82.6%)	51 (83.6%)	127 (83.0%)
Liver metastasis Percent missing	0 (0.0%)	0 (0.0%)	0 (0.0%)

Supplementary Table S7: Baseline characteristics of FFCD.

2023-5198 YODA DUA - Research proposal

Project Title

Evaluation of G-computation and conformal prediction to provide early signs of efficacy in single arm trial with time to event outcomes

Narrative Summary:

Phase II trials are often conducted using a single arm without comparative treatment and suffer from short follow-up times. Endpoints often reported in this context are the response measured using RECIST criteria based on the variation in the Sum of Longest Diameters (SLD) of the target lesions and the Prostate Specific Antigen (PSA) response in prostate cancer.

This study proposes to evaluate the use of conformal prediction to estimate for each patient the range of plausible variation in SLD and PSA level values that would have been observed under SoC, providing a point of comparison. Additionally, the use of G-computation will be explored based on the variation in SLD/PSA level.

Scientific Abstract:

Background : Phase II oncology trials are often single arm without comparative treatment and suffer from short follow-up times. An endpoint often reported and more suitable with short follow-up times is the response measured using RECIST criteria based on the variation in the Sum of Longest Diameters (SLD) of the target lesions at a given time. External control arm efficacy analysis based on this endpoint could be better powered and inform the decision to move to the next phase. Additionally, [Loiseau2022] suggests that G-computation increases statistical power compared to propensity based methodologies while controlling for type I error. Relying on G-computation to estimate treatment efficacy using change in SLD as endpoint could therefore be more informative in early phases. Conformal prediction emerged as a framework that allows for the construction of prediction intervals with guaranteed error bounds for a given outcome variable. Conformal prediction provides a measure of uncertainty around the predictions. In a setting of a phase II trial, conformal prediction could be used to predict for each patient what would have been the range of plausible change in SLD values whether he received comparative treatment. This could provide a point of comparison for each patient and potentially drive inclusion criteria of a phase III.

Objective :

- Evaluate G-computation directly applied to the largest reduction of the SLD and compare it with propensity score based estimators.
- Study the use of Conformal prediction to provide a point of comparison for each patient included in the single arm trial.

Study Design :

A pool of clinical trials that share a common treatment (Abiraterone acetate + prednisolone) will be used in this study. The outcomes of interest to compute the treatment effect will be the change in SLD and the PSA level.
To evaluate the relevance of relying on conformal prediction the following approach will be used. Given one trial A , we will consider all the patients under Abiraterone acetate + prednisolone in the pool of trials PVA, excluding the one considered, and restrict to the set of patients that share inclusion/exclusion criteria. All the patients in PVA will be used to derive a model to predict the change in SLD. Conformal prediction will then be used to produce intervals for the predictions that are guaranteed to contain the ground truth with 95 % probability. The model will then be applied on the Abiraterone acetate + prednisolone arm of the trial A to assess that the coverage of the methodology is as expected.
To evaluate the relevance of external control arm methodologies applied on SLD/PSA changes, and more particularly G-computation, we will rely on internal replication study [Loiseau2022].

Participants :

Individual data from all the trials specified.

Primary and Secondary Outcome Measure(s):

Change from baseline in SLD and in PSA level.

Statistical Analysis:

To assess the relevance of conformal prediction, we will compute the coverage (how many time the observed change in SLD/PSA falls into the predicted range of plausible values), the MSE, MAE and the width of the confidence interval.
To compare the different estimators of treatment effect on the change in SLD/PSA, we will compute the MSE, the MAE and the confidence interval width. We will also assess the ability of the methodology to reproduce the results of the original trial on hard endpoints.

Brief Project Background and Statement of Project Significance:

There is a growing interest in complementing a single arm with historical data. This is particularly true for Phase II trials conducted in oncology which often rely on a single arm testing the active treatment and lack of comparator. A white paper written by Medidata and FDA scientists was presented in a Friends of Cancer Research meeting in December 2018 and demonstrates the interest of both regulators and private companies in this question. Our work proposes to address one of the main limitations of this kind of methodology, the small number of events observed (progressions and deaths) in the single arm trials by providing analysis on an intermediate outcome (PSA/SLD change from baseline) and assess the relevance of this approach and could extend its use.
Additionally, we propose a more personalized approach to the external control arm by providing for each patient a range of plausible values under SoC relying on conformal prediction. This could open help selection of patients likely to maximize the benefit at an early stage.
Additionally, data is often spread and due to RGPD in Europe, pooling single arm trial data and real word data can be impossible. Therefore we evaluate the impact of relying on federated learning to deal with this limitation.

Specific Aims of the Project:

The objective of the project is to assess the relevance of the conformal prediction framework to provide an individual estimate of the counterfactual outcome, i.e. what would have been the range of plausible change in PSA/SLD values whether the patients received the SoC treatment instead of the treatment under assessment.
Additionally, we will assess the ability of G-computation in an ECA setting applied to change in PSA/SLD values to perform an estimation of efficacy in agreement with the estimate of the randomized trial on the hard endpoint.

What is your Study Design?

Methodological research

What is the purpose of the analysis being proposed?

Develop or refine statistical methods

Research on clinical trial methods

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

Given one trial A, we will consider all the patients under Abiraterone acetate + prednisolone in the pool of clinical trials PVA, excluding the one considered, and restrict to the set of patients that share similar background therapy and inclusion/exclusion criteria in order to comply with the positivity assumption required for causal inference.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

The study will focus on treatment effects on the change from baseline in SLD and change from baseline in PSA level.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

All the characteristics available at baseline will be considered for model training and validation.

Statistical Analysis Plan:

A pool of clinical trials that share a common treatment (Abiraterone acetate + prednisolone) will be used in this study. The outcomes of interest to compute the treatment effect will be the change in SLD and the PSA level. To evaluate the relevance of relying on conformal prediction to estimate the range of plausible change in SLD and PSA level, the following approach will be used.
Given one trial A, we will consider all the patients under Abiraterone acetate + prednisolone in the pool of clinical trials PVA, excluding the one considered, and restrict to the set of patients that share similar background therapy and inclusion/exclusion criteria in order to comply with the positivity assumption required for causal inference. All the patients in PVA will be used to derive a model to predict the change in SLD (or in PSA level) from baseline. Conformal prediction will then be used to produce intervals for the predictions that are guaranteed to contain the ground truth with 95 % probability. The model will then be applied on the Abiraterone acetate + prednisolone arm of the trial A to assess that the coverage of the methodology is as expected. On top of the coverage, we will also assess the MSE (mean squared difference between predicted and observed changes in SLD or PSA), MAE and the width of the confidence interval. Finally, we will apply the model to the other arm of the trial A and assess the number of times the predicted plausible SLD/PSA changes under Abiraterone acetate + prednisolone overlap with the observed outcome under the other treatment. We expect to observe a clear separation when the trial is successful.

To evaluate the relevance of external control arm methodologies applied on SLD/PSA changes, and more particularly G-computation, the following approach will be used. Given one trial A, we will consider all the patients under Abiraterone acetate + prednisolone in the pool of clinical trial PLA, excluding the one considered, and restrict to the set of patients that share similar background therapy and inclusion/exclusion criteria in order to comply with the positivity assumption required for causal inference. We will then perform the following experience:
Experiment 1: Assessing a treatment effect of zero between the Abiraterone acetate + prednisolone arm of A and the patients under Abiraterone acetate + prednisolone in PV;
Experiment 2: If A contains a comparator arm, we compare this comparator arm with patients under Abiraterone acetate + prednisolone in PVA. The confidence intervals obtained using G-computation on difference SLD/PSA changes are then compared to the confidence interval originally obtained using OLS and the two arms of the trials. Additionally the ability to recover the results on OS using the conclusion from the G-computation analysis.
To assess the relevance of the analysis relying on conformal prediction, we will compute the coverage (how many time the observed change in SLD/PSA falls into the predicted range of plausible values), we will also assess the MSE (mean squared difference between predicted and observed changes in SLD or PSA), MAE and the width of the confidence interval.
To compare the different estimators of treatment effect on the change from baseline in SLD and change from baseline in PSA level and assess the potential added value of G-computation, we will compute the MSE, the MAE and the confidence interval width. Additionally, when replacing the Abiraterone acetate + prednisolone arm of one trial by the set of patients that share similar background therapy and inclusion/exclusion criteria in other trials, we will assess the ability of the methodology to reproduce the results of the original trial on the hard endpoints (OS/PFS) using the regulatory agreement. The regulatory agreement is the percentage of time the cutoff P-value <0.05 obtained with the non-randomized experiments agrees with the RCT result about P-value <0.05. Software Used:
Python

Project Timeline:

Project start date: May 1, 2023 or when data accessed

Analysis completion: December 1, 2023
Manuscript draft completion: May 1, 2023

Dissemination Plan:

We plan on submitting this research as a research article in one of the following journals: 'Statistics in Medicine' or 'Statistical Methods in Medical Research' or 'BMS methodological research'.

Bibliography:

Loiseau, Nicolas, et al. "External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning." *BMC Medical Research Methodology* 22.1 (2022): 1-13.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107(5): 261--65. <https://doi.org/10.1257/aer.p20171038>.

Austin, Peter C., and Elizabeth A. Stuart. 2015. "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine* 34 (28): 3661 --79. <https://doi.org/10.1002/sim.6607>

Glynn, Adam N., and Kevin M. Quinn. 2010. "An Introduction to the Augmented Inverse Propensity Weighted Estimator." *Political Analysis* 18 (1): 36--56. <https://doi.org/10.1093/pan/mpp036>.

Robins, J. and A. Rotnitzky (1995). Semi-parametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90, 122-29.

Hahn, J. (1998). On the role of the propensity score in efficient semi-parametric estimation of average treatment effects. *Econometrica* 66, 315-31.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243\(263.

Andreux, Mathieu & Manoel, Andre & Menuet, Romuald & Saillard, Charlie & Simpson, Chloé. (2020). Federated Survival Analysis with Discrete-Time Cox Models.

Tibshirani, Ryan J., et al. "Conformal prediction under covariate shift." *Advances in neural information processing systems 32* (2019).

Barber, Rina Foygel, et al. "Conformal prediction under covariate shift." arXiv preprint arXiv:1904.06019 (2019).

Figure Supplementary Figure S5. YODA DUA Research Proposal. The research proposal submitted to the YODA DUA program.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RS.pdf