



Worm, D.T.H. (Daniël)¹; Kamphorst; B. (Bart)¹; Rooijackers, T.A. (Thomas)¹; Veugen, P.J.M. (Thijs)¹; Cellamare, M. (Matteo)²; Geleijnse, G. (Gijs)²; Knoors, D.J. (Daan)²; Martin, F. (Frank)²
¹ TNO; ² IKNL

Preserving privacy while analyzing vertically-partitioned cancer research data

Data from [Netherlands Cancer Registry \(NCR\)](#) often needs to be combined with other sources to include more information in studies about the prevention, care and outcomes of cancer. For example, next to cancer registry data, data on co-morbidities, medicine prescription or societal participation may be required.

Combining the data from different sources may compromise the privacy of the patients involved. In order to analyse such *vertically-partitioned* data in a privacy-preserving manner, we implemented solutions based on [Secure Multi-Party Computation \(MPC\)](#). MPC is a 'toolbox' of cryptographic techniques that enables several different parties to jointly analyze data, just as if they would have a shared database.

	time	event	attr X	
 Party 1	6	1	2	 Party 2
	7	0	0	
	4	1	1	
	2	1	3	
	⋮	⋮	⋮	

Vertically-partitioned data. Party 1 knows some attributes of his patients, party 2 knows complementary attributes of the same set of patients. MPC allows them to perform analyses on the combined data without disclosing data to each other.

Survival analysis is a particularly relevant type of analysis in cancer research. Two particularly often-used analyses are the Kaplan-Meier Estimator (KME), which is used to test whether two or more groups of patients significantly differ in terms of survival probability; and the Cox Proportional Hazards (CPH) model, which gives insight in characteristics that impact the chances of survival. Our goal is to run both types of analyses on distributed data in a privacy-friendly manner.

More of our privacy-preserving efforts:

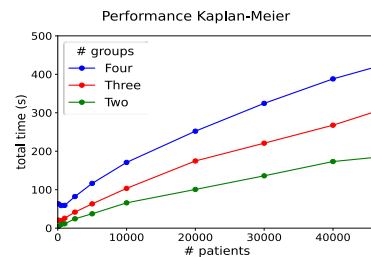
- CONVINCED project ([web](#), [report](#))
- Kaplan-Meier estimator ([poster](#), [report](#))
- Cox Proportional Hazard ([report](#))
- LASSO regression ([video](#))
- [Vantage6](#), the open-source Personal Health Train implementation



World-first: privacy-preserving log-rank test for the Kaplan-Meier Estimator

We designed and implemented* a fast and accurate privacy-preserving version of the log-rank test for the KME on vertically-partitioned data. The resulting p -value is accurate up to 5 digits, thus enabling the researcher to draw accurate conclusions for the tested hypothesis.

As seen in the plot below, the computation time for a total of 40.000 patients divided in four groups is in the order of minutes, with a sublinear scaling in the number of patients.



World-first: privacy-preserving Cox Proportional Hazards model training

Continuing our efforts to expand the privacy-preserving survival analysis toolbox, we designed and implemented* a privacy-preserving version for training the CPH model. Several important building blocks, including a secure exponentiation function, had to be designed in the process. The resulting privacy-preserving CPH model shows accurate results on our test data, but more research is required in order to make it sufficiently fast and scalable for practical use.

Conclusions and Future Work

Cryptographic innovations enable development of privacy-preserving survival analyses on vertically-partitioned data. We successfully implemented a sufficiently accurate and fast privacy-preserving version of the log-rank test for KME and developed a demanding, yet promising basis for privacy-preserving CPH model training. These will be integrated into [Vantage6](#), the open-source Personal Health Train implementation. Within the TKI HTSM project LANCELOT and Appl.AI project SELECTED, TNO and IKNL joined by new partners, continue research on privacy-preserving analyses on vertically-partitioned data.

*In both solutions, the open source Python library [MPyC](#) (based on Shamir Secret Sharing). The KME solution also made use of the Paillier cryptosystem (homomorphic encryption).