Authors: Passey A, Perualila, N., Verbeke T., Nassiri V., Van Speybroeck M

Title: HONEUR (Haematology Outcomes Network in Europe) – Distributed statistics in a Federated Model to Support Real World Data Research in Hematology

Introduction: The Haematology Outcomes Network in EURope (HONEUR) is an interdisciplinary initiative aimed at improving patient outcomes by analyzing real world data across hematological centers in Europe. Its overarching goal is to create a secure network which facilitates the development of a collaborative research community and allows access to big data tools for analysis of the data. The central paradigm in the HONEUR network is a federated model whereby the data stays at the respective sites and the analysis is executed at the local data sources. To allow for a uniform data analysis, the common data model 'OMOP' (Observational Medical Outcomes Partnership) was selected and extended to accommodate specific hematology data elements. While the federated model addresses ethico-legal challenges for data pooling, it poses specific challenges when performing statistical analysis on pooled data.

Objective: Enabling the use of distributed statistics across federated data models

Methods: To validate the feasibility and accuracy of distributed statistics in the HONEUR network, data from the EMMOS registry (NCT01241396) were used. This registry is a prospective, non-interventional study that was designed to capture real world data regarding treatments and outcomes for multiple myeloma at different stages of the disease. Data was collected between Oct 2010 and Nov 2014 on more than 2,400 patients across 266 sites in 22 countries. After mapping data to the OMOP common data model version 5.3, the three most populated countries in the dataset: Germany, Italy and Russia with 363, 488 and 213 subjects, respectively, were selected. Therefore, a total of 1064 patients comprised the pooled dataset. In this work, the overall survival of the patients is modeled with age, sex and Salmon-Durie stage as model covariates. Two types of analysis were performed: (1) the traditional Cox regression model stratified by country using the pooled dataset and (2) the same cox regression model distributed by country. The second analysis allows for estimating model parameters without the need for countries to share patient-level data. This distributed model is based on the R *distcomp* package from Narasimhan *et al*., which employs homomorphic computation and allows for the calculation of the overall likelihood function across study sites and estimation of model parameters.

Results: Using real world patient haematology data we were able to demonstrate identical results from the two analyses performed using the pooled dataset versus distributed dataset(s) as shown in Tables 1 and 2, respectively. The hazard ratios, 95% confidence intervals and p-values are identical between the two models for all for levels of stage compared to reference, adjusted for age and gender indicating that the distributed model has generated precisely the same proportional hazard estimate as well as variance estimates for variables within the model.

Conclusions: We have compared a standard pooled approach for a comparative outcomes study using available clinically relevant variables with a distributed analysis using the same parameters where the analyst has no access to the patient level data, the methodology has generated a virtually identical underlying fitted model with the precisely the same effect estimates for hazard, confidence intervals and p-values from hypothesis testing indicating that this model has great potential for testing more complex comparative effectiveness modelling with accurate outputs which entirely protect patient privacy in a federated data model.

Table 1: Coefficients from pooled model

	Covariate	HR	Lower_CI_95	Upper_CI_95	P_value
1	Stage 1 (REF)	1.00	NA	NA	NA
2	Stage 2	0.83	0.52	1.33	0.449
3	Stage 3	1.19	0.80	1.77	0.398
4	Gender M v F	1.05	1.04	1.07	0.000
5	Age (linear)	1.13	0.86	1.49	0.366

Table 2: Coefficients from distributed model

	Covariate	HR	Lower_CI_95	Upper_CI_95	P_value
1	Stage 1 (REF)	1.00	NA	NA	NA
2	Stage 2	0.83	0.52	1.33	0.449
3	Stage 3	1.19	0.80	1.77	0.398
4	Gender M v F	1.05	1.04	1.07	0.000
5	Age (linear)	1.13	0.86	1.49	0.366