MSR57

Survival analysis without sharing of individual person data: an antidote to "data available upon request"?

FEDERICO BONOFIGLIO, VERAMED GERMANY
federico.bonofiglio@veramed.co.uk

Introduction

Cox regression and Kaplan-Meier (KM) estimates are often needed in post-approval clinical research which requires access to individual person data (IPD). For example, a long-term treatment effect review or a comparative effectiveness oncology study involving the IPD owned by a competitor. Here, IPD cannot always be shared due to privacy or proprietary restrictions.

Methods

I propose a method that generates pseudodata for analysis by replacing the IPD with non-disclosive IPD aggregates (Moments) to be shared with a central computer as input parameters to a Gaussian copula (GC). This generates the pseudodata, allowing analysts to access completely anonymised versions of possibly multiple datasets on which to run analyses. **Survival inferences are computed on the pseudodata as if it was the IPD.**

IPD Moments → Gaussian Copula → Pseudodata → Analysis

Results

Perhaps counter-intuitively, IPD Moments can yield enough information to approximate KM and Cox regression results. Other generative methods are available like Classification and Regression Trees (CART) which require IPD as input data. GC estimates are comparable to CART, although GC might not perform as well for sub-group analyses (not shown).

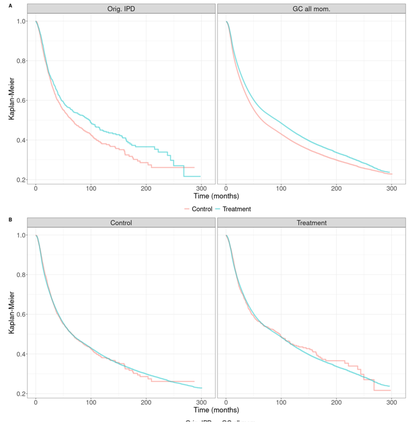
GC avoids many legal problems related to IPD privacy or property while facilitating approximation of common IPD survival analyses that are otherwise difficult to conduct.

Standard publication of IPD Moments could facilitate "second purpose"-research activities.

Survival Analysis without Individual Person Data

veramed
End-to-End Biometrics

Application to the GASTRIC study: 14 centers, a total of 3288 patients, adjuvant chemotherapy (treatment) vs. surgery (control). Each center only shares IPD Moments. Centrally, GC generates pseudodata and produces KM and Cox-regression estimates.

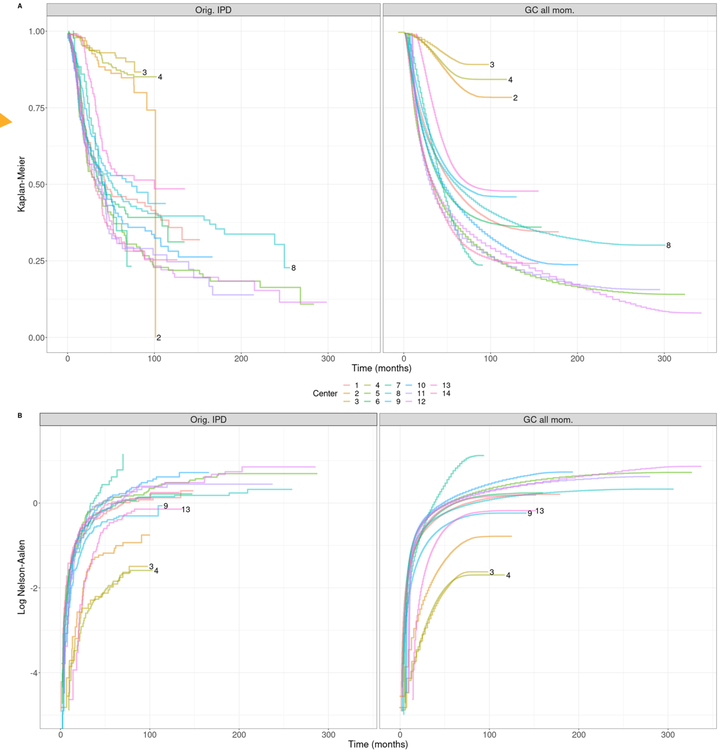


KM estimates made on original IPD and pseudodata are comparable.

Several Cox regression models and tests can be reproduced using the pseudodata. In a multi-center setting, using the anonymous pseudodata yields more information than summary based meta-analysis and akin to using the original IPD.

			IPD non disclosive & summary based		
			Gaussian copula		
Inference	Orig. IPD	Dist. Boot	Fewer mom.	All mom.	Meta an.
Hazard Ratio (SD) Wald test (HR)	-0.153 (0.049) -3.139 (0.002)	-0.152 (0.049) -3.104 (0.002)	-0.145 (0.049) **-2.979** (0.003)	-0.154 (0.049) -3.151 (0.002)	-0.151 (0.049) -3.1 (0.002)
PH test (Wald)	-0.016 (0.987)	**-1.255 (0.21)**	-0.041 (0.967)	**-0.986 (0.324)**	-0.008 (0.994)
Random Center Effect (χ test)	666.813 (0)	**677.21** (0)	**729.156** (0)	**679.904** (0)	· (·)
Median Hazard Ratio	0.751 (2.047)	0.758 (2.06)	0.782 (2.108)	0.761 (2.068)	· (·)
Random treatment Effect χ test	0.003 -0.002 (1)	**0.132** **2.321 (0.128)**	**0.114** **1.861 (0.173)**	**0.114** **2.96 (0.085)**	**0.153** 11.43 **(0.575)**
Random effects ρ (MSE)	0.999 (·)	**0.438** (0.44)	**0.347** (0.477)	**0.177** (0.481)	· (·)

Center-specific estimates: See successful KM anonymization for Center 2.



Legend

Orig. IPD = *Original IPD.*
GC all mom. = *pseudodata generated by GC using means, SDs, skewness, kurtosis and pairwise Pearson correlations.*
GC fewer mom. *= excludes skewness and kurtosis.*
Meta an. *= Summary based Meta analysis*
Dist. Boot. *= Bootstrap of original IPD distributed across centers.*
PH *= Proportional Hazards.*
P *= Random effects correlation.*
MSE *= Mean Squared Error.*

Notes

Result diverging from Orig. IPD by at least 0.05 in absolute value is boldfaced (except for Meta an. χ test which refers to the heterogeneity Q-test).

Scan the QR Code with your phone to view the full paper