INFERENCE IN CURE MODELS WHEN THE FOLLOW-UP IS INSUFFICIENT

A Data Management Plan created using DMPonline.be

Creator: Ingrid Van Keilegom

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: G047524N

ID: 208504

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

In traditional survival analysis, all subjects in the population are assumed to be susceptible to the event of interest. However, it often happens that the event of interest never occurs for a fraction of the subjects under study, that are said to be cured. The quantities of interest in these cure models are often the cure rate and the survival function of the susceptible or uncured individuals. When the survival time is subject to random right censoring, as is common in survival analysis, all cured subjects will be censored, whereas the non-cured ones can be either censored or uncensored. Hence, in order to identify the cure fraction we need to impose certain assumptions on the model. A common way to identify a cure model is to impose the so-called assumption of sufficient follow-up. When the follow-up period is erroneously believed to be sufficient, the cure rate will be overestimated leading to possibly false (and too positive) conclusions. Therefore, it is important to have methods at hand that are able to correctly estimate the cure rate and the distribution of the uncured. The current literature on estimation and testing under insufficient follow-up is still in a very early stage of development, and much more work is needed to do correct inference in this important research area. Our goal is to propose novel estimation methods that overcome the problem of insufficient follow-up, and powerful testing strategies for the hypothesis of sufficient follow-up.

Last modified: 24-06-2024

INFERENCE IN CURE MODELS WHEN THE FOLLOW-UP IS INSUFFICIENT
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
(JTPA)	The National Job Training Partnership Act Study	reused	digital	observational	DTA,SD2 and CSV	<1GB	
R scripts	Code written to apply new methods to data	generate	digital	software	.R scripts	estimated <100MB	
Numerical results	results from simulations	generate	digital	simulation data	text (.txt) and Excel (.xlsx) files	estimated around 5GB	
Manuscripts	text to describe new methods, methodologies and results	generate	digital	other	stored on Overleaf and on my laptop as .tex files and PDF files	estimated to be around 10GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

JTPA: https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

R-scripts:
Every R file will have some comments added to the script.
Numerical results:
All the generated data sets and numerical results are stored locally, with the file name including the time, date and specific setting it was generated from.
Manuscripts:
They are stored locally on my laptop and on Overleaf. Back-ups are made from time to time so that there is also some version control.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

All data will be stored locally on my laptop (protected by a password).

How will the data be backed up?

Backups are made on a secure hard drive every week. In addition all files on the laptop are also backed up on iCloud.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

I will not need more capacity than what is available on iCloud. Extra storage can always be bought when necessary for hard drive backups.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data on iCloud are password protected. The hard drive with the backup is safely kept at home.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

No costs are expected.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 15 years.

Where will these data be archived (stored and curated for the long-term)?

All data will be preserved on my laptop (or new laptops when the current one will need to be replaced), on the external hard drive and on iCloud. The JTPA data is publicly available.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No costs are expected.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

All finished manuscripts will be made publicly available on arXiv. Some of the R-scripts and numerical results will be available on the Github pages of my co-authors. R-scripts and numerical results that cannot be found on these Github pages, will be available to interested researchers upon request via email. The JTPA data is publicly available.

If access is restricted, please specify who will be able to access the data and under what conditions.

//

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All finished manuscripts will be made publicly available on arXiv. Some of the R-scripts and numerical results will be available on the Github pages of my co-authors. R-scripts and numerical results that cannot be found on these Github pages, will be available to interested researchers upon request via email. The JTPA data is publicly available at https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study.

When will the data be made available?

When manuscripts are submitted for publication.

Which data usage licenses are you going to provide? If none, please explain why.

None.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

No costs are expected.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Ingrid Van Keilegom

Who will manage data storage and backup during the research project?

Ingrid Van Keilegom

Who will manage data preservation and sharing?

Ingrid Van Keilegom

Who will update and implement this DMP?

Ingrid Van Keilegom