STRUCTURED OUTPUT LEARNING FOR HEALTHCARE APPLICATIONS

A Data Management Plan created using DMPonline.be

Creator: n.n. n.n.

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: G046024N

ID: 208445

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

Machine learning has great application potential in healthcare. Yet many healthcare applications have data characteristics that make it difficult to apply standard machine learning methods. For example, patient outcomes can often be expressed in different ways and are difficult to summarize in one value. This requires the application of so-called structured output learning methods. Such methods are not yet widely used in practice, because it is often difficult to find an exact mapping from the applications to the methods currently available in the literature. In this project we want to fill this gap by contributing to the structured output learning domain. More precisely, we will design novel multi-task learning, multi-instance learning and label ranking algorithms. Our algorithmic contributions are motivated by two concrete clinical applications, concerning paediatric intensive care unit patients and infertile couples. We will be working with real data from these applications. On the longer term, our methods will assist clinicians in providing personalised and preventive healthcare.

Last modified: 25-06-2024

STRUCTURED OUTPUT LEARNING FOR HEALTHCARE APPLICATIONS
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	
Fertility dataset (ART)	medical information (patient records and embryo characteristics) for patients with an assisted reproductive technology (ART) treatement.	Reuse existing data	Digital	Observational	.csv	<1GB	
Intensive Care dataset (ICU)	medical information (patient records) for pediatric patients in the intensive care unit, complemented with neurocogntive outcomes assessed afterwards.	Reuse existing data	Digital	Observational	.csv	<1GB	

Processed data	For both above data sources, the data will be cleaned and intermediate features will be defined, making the data ready to be used for modelling	Generate new data	Digital	Compiled/Aggregated data	.csv	<1GB	
Computer programming code	Computer programming code (data analysis, machine learning algorithms)	Generate new data	Digital	Software	.py, .R	<100MB	
Analysed data	Machine learning models, graphs, tables, text	Generate new data	Digital	Compiled/Aggregated data	.pdf, .txt	<1GB	
Research papers	Manuscripts to be published	Generate new data	Digital	Other	.pdf, .tex, .docx	<100MB	

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:

ART: Source: UZ Leuven database, data to be extracted
ICU: Source: data collected in Pepanic and TGC follow-up studies
Associated publications:
DOI: 10.1016/S2213-2600(18)30334-5
DOI: 10.1016/S2352-4642(20)30104-8
DOI: 10.1001/jama.2012.12424

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.

- Yes, human subject data

ART:
Ethical approval to be obtained later in the project
ICU:
Covered by earlier ethical approval:
S54127 (ML8052): Impact of early parenteral nutrition completing enteral nutrition in pediatric critically ill patients, PI Prof. Dr. Greet Van den Berghe
ML2586: Tight glycemic control with intensive insulin therapy in PICU, PI Prof. Dr. Greet Van den Berghe

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.

- Yes

We will process pseudonymized retrospective data from ART and ICU patients followed up at UZ Leuven. This involves demographical characteristics, measurements taken at hospital visits, embryo related characteristics (ART data), and neurocognitive characteristics obtained at

follow-up (ICU data).

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.

- Yes

The developed software may result in valorisation potential, e.g. for companies developing clinical decision support tools for use in ART or ICU context. In that case, the exact valorisation path that will be used should be discussed with the involved research teams and LRD and will be updated later.

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).

Documentation will include information on how to pre-process the data (inclusion/exclusion criteria for patients, which variables to select, how to deal with missing data, etc.), as well as the exact validation procedure (train/test split, cross validation, evaluation measures) and parameter optimization procedure. Although a detailed methodology section will be required in any publication about the project, we will have a more elaborate description of the exact protocol in an extra document, this may be a readme.txt file or a lab notebook. For the generated computer programs, detailed source code documentation and a manual will be added.

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

The metadata of the generated software includes programming language, author, version, date of creation,... Such metadata is automatically created by a git repository like KU Leuven GitLab.

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

Where will the data be stored?

UZ Leuven datasets (ART and ICU) will be stored on a shared network drive managed by KU Leuven ICTS or on OneDrive for Business cloud storage with multifactor authentication with the KU Leuven Authenticator.
Processed data will be stored in the same locations as the raw data.
Analysed data will be stored on the PhD student's hard drive.
Programming Code will be stored on the PhD student's hard drive and KU Leuven GitLab. The code may be made available upon publication of an article describing the developed method.
Publication data will be stored on the PhD student's hard drive and in Lirias.

How will the data be backed up?

We will use standard back-up provided by KU Leuven ICTS.

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

Hard drives in our group have a capacity of at least 500 GB, which is more than sufficient to store all the data. Our OneDrive for Business license offers 2TB of storage. Furthermore, KU Leuven GitLab offers sufficient storage for the double backup of the code generated in this project.

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

Servers in Leuven (for the UZLeuven data): responsibility of the UZ Leuven team. Hard drive of our laptops: has a bitlocker, laptop itself is password-protected. OneDrive for Businesss: secured by two-factor authentication.
KU Leuven GitLab: secured by two-factor authentication.
The latter 2 storage media are supported by the KU Leuven infrastructure, and are ensured to stay within a datacenter in Europe.

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

No substantial costs are expected, as our data volumes will stay well below the standard limits.

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 generated data will be retained for minimum 10 years (KU Leuven RDM policy).
The preservation of the reused patient data falls under the responsibility of the UZ Leuven teams.

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

The generated software data will be archived on KU Leuven's K drive, which is a drive specifically for archiving. It may also be made available, e.g. on a GitHub repository.

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

The expected costs for preservation are negligible and will be covered by the budget of our research group.

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

Raw data reused in this project and processed data will not be made available for legal and ethical reasons. The processed data can easily be recovered from the raw data using the software code generated during this project.
Analysed data will be included in a publication, possibly as supplemental materials.
Code may be made available, e.g. in a hosted GitHub repository. Access to these repositories will not be restricted, as they do not consist of sensitive personal data.

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

Not applicable.

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.

- Yes, Privacy aspects
- Yes, Ethical aspects

We are not allowed to (re)share the UZ Leuven data that is being reused here and the processed data derived from it.

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

In an Open Access repository.

When will the data be made available?

Upon publication of the research results.

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

The exact license of the software has to be decided, and will be discussed with LRD.

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

The generated data will be linked to a publication, which by default will have an associated DOI.

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

No costs for data sharing are expected, creation of a public GitHub repository is free.

6. Responsibilities

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

PhD student(s) / postdoc hired on this project.

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

PhD student(s) / postdoc hired on this project.

Who will manage data preservation and sharing?

PhD student(s) / postdoc hired on this project together with PI (Celine Vens)

Who will update and implement this DMP?

PhD student(s) / postdoc hired on this project together with PI (Celine Vens)