Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241096

PROSurvival: A Technical Case Report on Creating and Publishing a Dataset for Federated Learning on Survival Prediction of Prostate Cancer Patients

Tingyan XU[a,1], Timo WOLTERS[a], Johannes LOTZ[b], Tom BISSON[c], Tim-Rasmus KIEHL[c], Nadine FLINNER[d], Norman ZERBE[c,e] and Marco EICHELBERG[a]
[a] *R&D Division Health, OFFIS - Institute for Information Technology, Germany*
[b] *Fraunhofer Institute for Digital Medicine MEVIS, Germany*
[c] *Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institut für Pathologie, Berlin, Germany*
[d] *Goethe University Frankfurt, Universitätsklinikum, Dr. Senckenbergisches Institut für Pathologie, Frankfurt am Main, Germany*
[e] *Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institut für Medizinische Informatik, Berlin, Germany*
ORCiD ID: T Xu https://orcid.org/0009-0005-8791-9582, J Lotz https://orcid.org/0000-0003-3387-2596, T Bisson https://orcid.org/0000-0002-7743-0792, TR Kiehl https://orcid.org/0000-0003-0616-7082, N Flinner https://orcid.org/0000-0002-8565-6492, N Zerbe https://orcid.org/0000-0002-0314-3037, M Eichelberg https://orcid.org/0000-0002-8590-3318

Abstract. The PROSurvival project aims to improve the prediction of recurrence-free survival in prostate cancer by applying federated machine learning to whole slide images combined with selected clinical data. Both the image and clinical data will be aggregated into an anonymized dataset compliant with the General Data Protection Regulation and published under the principles of findable, accessible, interoperable, and reusable data. The DICOM standard will be used for the image data. For the accompanying clinical data, a human-readable, compact and flexible standard is yet to be defined. From the set of existing standards, mostly extendable with varying degrees of modifications, we chose oBDS as a starting point and modified it to include missing data points and to remove mandatory items not applicable to our dataset. Clinical and survival data from clinic-specific spreadsheets were converted into this modified standard, ensuring on-site data privacy during processing. For publication of the dataset, both image and clinical data are anonymized using established methods. The key challenges arose during the clinical data anonymization and in identifying research repositories meeting all of our requirements. Each clinic had to coordinate the publication with their responsible data protection officers, requiring different approval processes due to the individual states' differing interpretations of the legal regulations. The newly established German Health Data Utilization Act is expected to simplify future data sharing in a responsible and powerful way.

[1] Corresponding Author: Tingyan Xu, OFFIS-Institute for Information Technology, Escherweg 2, 26121 Oldenburg, Germany; E-mail: tingyan.xu@offis.de.

Keywords. Research data, whole-slide images, de-identification, DICOM, GDPR, FAIR, oBDS

1. Introduction

During cancer diagnosis and treatment, significant clinical data are collected. This includes tissue samples from biopsies and radical prostatectomy specimens from patients with prostate cancer (PCa). Histological glass slides are produced form these samples for microscopic diagnosis by pathologists. These slides can be digitized into whole-slide images (WSIs), which in turn can be used to train machine learning models to automate diagnosis, discover biomarkers, and improve prognosis. In the PROSurvival project, this data is used to train a neural network for the prediction of clinical outcome, such as recurrence, using federated learning. Furthermore, the training dataset will be made available to the research community under the FAIR (Findable, Accessible, Interoperable, Re-usable) principles. For this purpose, clinical data must be standardized to ensure interoperability. WSIs may contain sensitive information, requiring a thorough anonymization [1] as well as standardization, for which we chose DICOM (Digital Imaging and Communications in Medicine) [2].

Standardization of clinical data can basically be considered a solved problem; therefore, this paper discusses the adaptation of a standard format for PCa patient data, comprised of the following items:

- **ID**: A dataset-specific identifier linking clinical and image data.
- **TNM classification**: The post-surgical histopathological classification, especially the pathological TNM (pTNM), consisting of the following components: **T** (Tumor) indicates the size and extent of the primary tumor, **N** (Nodes) describes the absence or presence and extent of regional lymph node involvement, and **M** (Metastasis) denotes cancer spread to other parts of the body. Additional descriptors are **L** for lymphatic vessel, **V** for venous and **Pn** for perineural invasion
- **R-Status**: Resection margin status, indicating whether the tumor is present at the surgical edge, implying possible need for further surgery.
- **PSA**: Prostate-Specific Antigen levels measured in blood; high levels can indicate PCa presence and correlate with tumor size. Post-prostatectomy, PSA should be undetectable in a completely resected tumor without metastasis. A rising PSA value may indicate recurrence.
- **Gleason Score:** Grades assigned based on the histologic appearance of PCa. The **primary** score denotes the **most common** tumor pattern in the sample and the **secondary** score denotes the **worst** or most abnormal gland structure observed. A **tertiary** score may be included for the second most common cell structure.
- **ISUP score:** Grouped grades developed by the International Society of Urological Pathology (ISUP), providing a simplified and standardized classification.
- **Survival Times:** Overall Survival (**OS**) is calculated from PCa diagnosis to death; Disease-Specific Survival (**DSS**) equals OS if death is due to PCa; Recurrence-Free Survival (**RFS**) is tracked via PSA levels post-prostatectomy, where a PSA level higher than 0.2 ng/ml indicates recurrence.
- **Clinical trajectory** data, such as PSA levels measured during post-prostatectomy follow-up visits, or date and cause of death. This follow-up data is crucial for training machine learning models. We selected data from patients who underwent prostatectomy in 2013-2018, ensuring availability of long-term follow-up data. The

dataset, however, is incomplete as not all patients were followed up at the same hospital. Therefore, the dataset must accommodate this incompleteness while remaining human-readable and aligned with existing standards.

There are several published datasets on prostate cancer (PCa) or other cancer cases with similar imaging and clinical data [3,4]. For instance, **Zhong Q et al.** [5] published a curated set of tissue microarray (TMA) images and clinical outcome data for prostate cancer patients in 2017. Their extensive clinical data is provided in a spreadsheet in SPSS format. Additionally, **the Cancer Genome Atlas (TCGA)** [6] offers datasets for prostate and other cancers, with clinical data available in non-standardized TSV or JSON formats.

Existing standards for sharing clinical data on cancer patients include formats such as **OMOP** [7], which is primarily used in clinical trials and research. OMOP provides a standard with a defined vocabulary and extensive information, much of which is irrelevant to PROSurvival and lacks defined semantics in the data model. The German Medical Informatics Initiative (**MII**) [8] has developed a core dataset standard in FHIR (Fast Healthcare Interoperability Resources) format, including a dataset for pathological diagnoses and results. This standard was still under development during the conceptualization period of PROSurvival and therefore excluded from the set of eligible standards. The **oBDS** (Bundeseinheitlicher onkologischer Basisdatensatz, Uniform National Oncological Basic Dataset) [3], an XML format used to report clinical data in a standardized form to German cancer registries, is well-defined and familiar to healthcare professionals but is more complex than our project requires. However, it is flexible enough to be modified and extended to suit our specific requirements.

2. Methods

After reviewing existing standards, we chose to extend the oBDS format to align with the specific needs of our project. Originally, oBDS was developed for reporting cancer patients' diagnoses and treatments to cancer registries, with a focus on rigid, mandatory data elements within an event-based reporting schema. To better serve our purposes, we modified the format by removing unnecessary components, such as patient name and address, as well as specific dates, such as the date of diagnosis. We also transitioned from the event-based structure to a patient-centered format, where multiple events are consolidated into a single report or file. Since the oBDS format is based on XML and described by a single XSD document, it allowed us the flexibility to extend, reduce, and tailor the format to our requirements. Additionally, to adhere to the FAIR principles and facilitate international data sharing, we translated the format from German into English. Key modifications included:

- **Identify Required Data Fields**: Non-relevant fields and modules for other cancer types were removed.
- **Field Existence Checks**: All fields, except the ID, were made optional to account for incomplete follow-up data.
- **Gleason Score:** We added a third Gleason score field to document the second most common cell structure and included a pathologist ID field for score differentiation. Gleason scores for biopsy and prostatectomy samples were better documented by moving the reason field inside the score structure. The ISUP field was also incorporated within the score structure.

- **PSA:** PSA values were categorized into *<10 ng/ml* and *>=10 ng/ml* as part of the anonymization process.
- **Survival time calculations**: Survival times were defined with specific fields for existence and duration (in months, up to 200).

Raw data was manually extracted by pathologists from local hospital information systems, stored in spreadsheets, and then locally processed using a Python script, yielding anonymized XML files for each patient. The script parsed the spreadsheets, performed necessary calculations, and generated corresponding XML files.

3. Results

Following the successful transformation based on an anonymized subset of data from both clinics, the script was then locally applied to the complete data from one of the clinics. The data, which had been collected but not standardized, presented several challenges due to error-prone nature of the manual information extraction process. For example, the TNM classifications were stored in different formats and sometimes lacked all classification markers. Similarly, Gleason scores included variations, such as percentages denoting specific cell structures.

Another significant adjustment involved handling the dates of follow-up PSA values. These dates were often recorded imprecisely, sometimes only stating the month or even the year, requiring the script to accommodate such vagueness.

After adapting the script to manage these exceptions, the clinical data was successfully converted into XML format and fed into the deep learning algorithm for analysis. Additionally, XML files containing a subset of the survival data were created for publication, ensuring compliance with privacy standards. We have also identified the need for further adjustments, specifically regarding the calculation of Recurrence-Free Survival (RFS). Moving forward, if there is no known recurrence, the RFS event will be set to 0, and the time will be calculated accordingly using the date of last contact with the patient. Conversely, if the PSA does not decrease to the nadir, here defined as 0.2 ng/ml, following treatment, the RFS will be marked as 1, with the time set to 0. These refinements will allow for more accurate and consistent tracking of RFS within the dataset.

4. Discussion

The data collection process was both challenging and time-consuming, primarily due to the differences in the hospital information systems used by the two clinics. These systems vary in architecture and data organization, with information distributed across multiple departments, each with distinct responsibilities. While this setup works well for internal communication and patient care, it significantly complicates data extraction for research purposes. Some of the data, over ten years old, was unsorted and required expert interpretation by pathologists to ensure accuracy.

Compliance with the General Data Protection Regulation (GDPR) added another layer of complexity. The interpretation of GDPR requirements varies across German federal states, necessitating extensive discussions with data protection officers at each location. Achieving a consensus on a reduced dataset for publication required careful

negotiations between the clinics and their respective data protection authorities to balance data privacy with the need for meaningful research.

Accessing valuable data from cancer registries also proved difficult due to lengthy processing times and various restrictions. Additionally, as some epidemiological cancer registries in Germany only became operational after 2013, relevant data may not have been fully reported.

Finding a suitable repository for research data presented another challenge. Whole-slide images (WSIs) are large and require specialized repositories, many of which do not support the storage of accompanying clinical data. To address this, we partnered with BBMRI-ERIC, the European research infrastructure for biobanking, to host our data.

5. Conclusions

This paper details the process of modifying the oBDS format to meet the specific needs of the PROSurvival project, which focuses on the survival of prostate cancer patients. The project encountered significant challenges, particularly in data collection and ensuring GDPR compliance. The differences in hospital information systems and the varying interpretations of GDPR across German states required careful adjustments and negotiations to maintain both data privacy and research integrity. A key challenge was finding a suitable repository for storing and sharing clinical and image data, which we successfully addressed by partnering with BBMRI-ERIC.

Despite these hurdles, the modifications to the oBDS format have laid a strong foundation for creating a standardized and anonymized dataset, ideal for advanced machine learning applications. While navigating complex regulatory landscapes, this project has also paved the way for future research to benefit from the recently established German Health Data Utilization Act (GNDG), which promises to simplify data sharing and utilization by providing a clearer legal framework.

Acknowledgement: The PROSurvival project is funded by the German Ministry of Research and Education, grant no. 01KD2213.

References

[1] Bisson T, et al. Anonymization of whole slide images in histopathology for research and education. Digit Health. 2023 May 9; 9:20552076231171475. doi: 10.1177/20552076231171475.
[2] DICOM Standard Definition 2024c [Internet]. The Medical Imaging Technology Association; 2024 [cited 2024 Sep 30]. Available from https://www.dicomstandard.org/current
[3] oBDS - Uniform national oncological basic dataset [Internet]. 2024 [cited 2024 Sep 30]. Available from https://basisdatensatz.de/basisdatensatz
[4] Gesundheitsdatennutzungsgesetz [Internet]: Bundesministerium der Justiz; 2024 [cited 2024 Sep 30]. Available from https://www.recht.bund.de/bgbl/1/2024/102/VO.html
[5] Zhong Q, et al. Data Descriptor: A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients. 2017 Jan; Scientific Data 4:170014; DOI: 10.10.1038/sdata.2017.14
[6] The Cancer Genome Atlas program [Internet]. United States of America: National Cancer Institute; 2024 [cited 2024 Sep 30]. Available from: https://www.cancer.gov/ccg/research/genome-sequencing/tcga
[7] Standardized Data: The OMOP Common Data Model [Internet]. 2024 [cited 2024 Sep 30]. Available from: https://www.ohdsi.org/data-standardization/
[8] MII Core dataset pathology module. 2024 [cited 2024 Sep 30]. Available from: https://www.medizininformatik-initiative.de/Kerndatensatz/Modul_Pathologie_Befund/MII-IG-KDS-Modul-Pathologie-Befund-Index.html