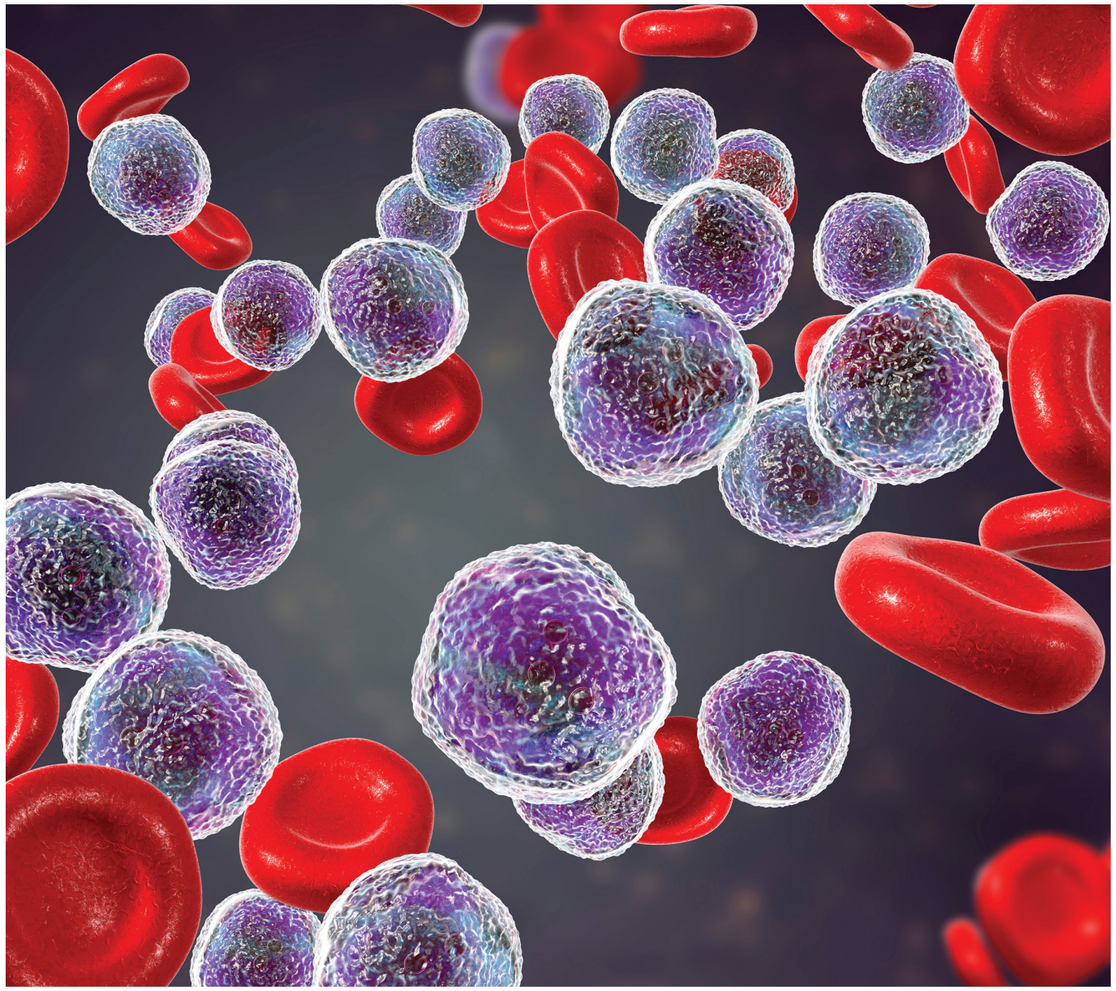
Copenhagen Ultrathon on Precision Medicine

Carsten U. Niemann
Rudi Agius

CLL-TIM

Challenge ID: U21-06



Main research question: Predict risk of infection (blood culture drawn) and chance of treatment free survival 4 years from start of first CLL treatment.

Secondary research question(s): Group patients according to phenotype (unsupervised clustering) and detect outcomes correlated with these groups. Develop meta learning approaches to adapt and validate these findings to other diseases/ cohorts.

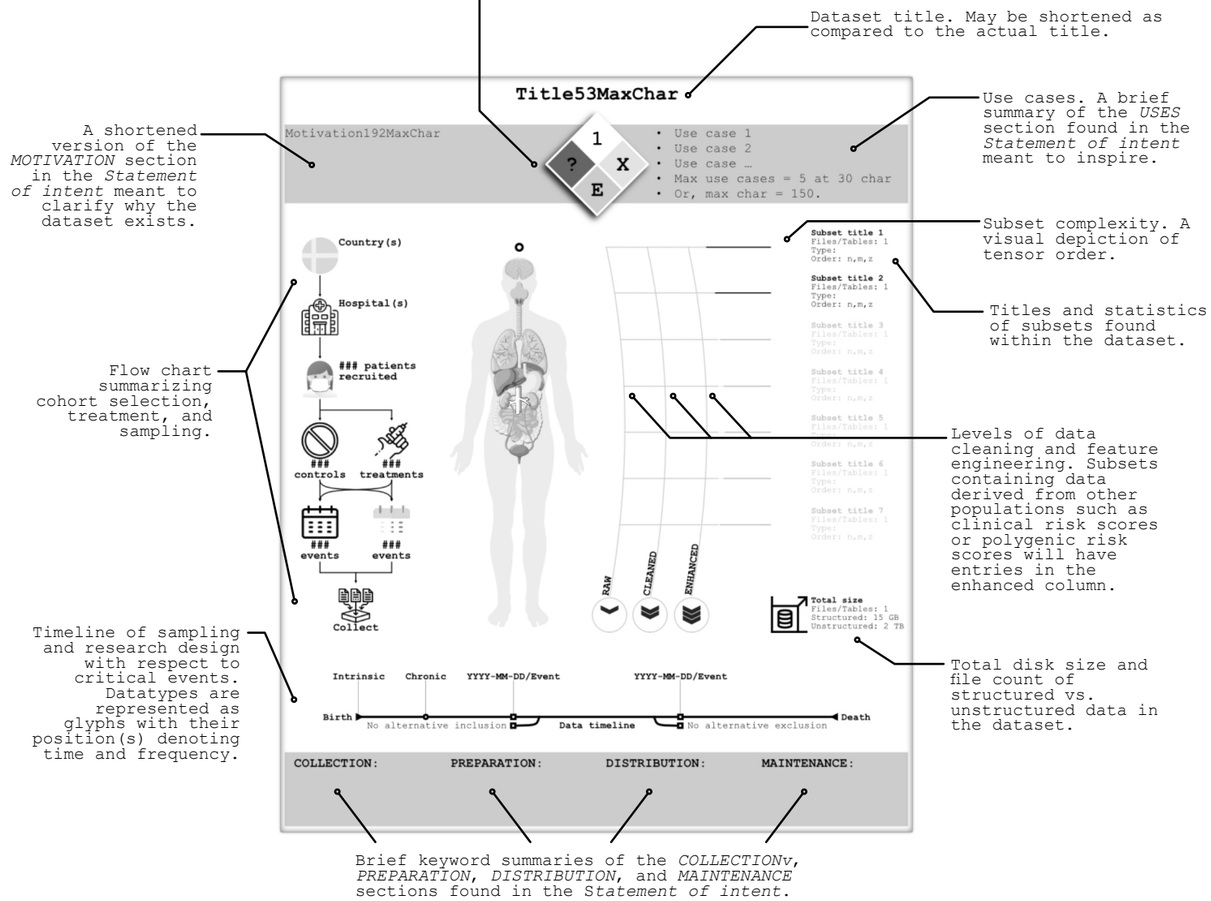
KEY

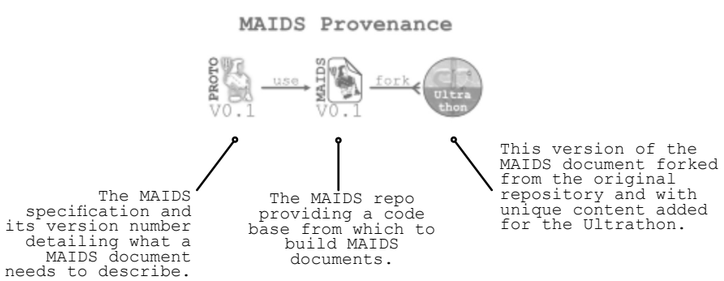
Document score from 1 to 3, 1 being basic and 3 being excellent.

Synthetic score from E to A, E having no synthetic data and A having synthetic data equivelant to the original.

Sample size in orders of magnitude {X¹, X², X³, etc}.

Independent evaluation of bias and fairness in the dataset with respect to the population it was created to represent. Symbols mean:
- "?" No evaluation has taken place.
- "!" Extreme levels of bias.
- "✓" Almost no bias found.
- "~" A mix, be wary.





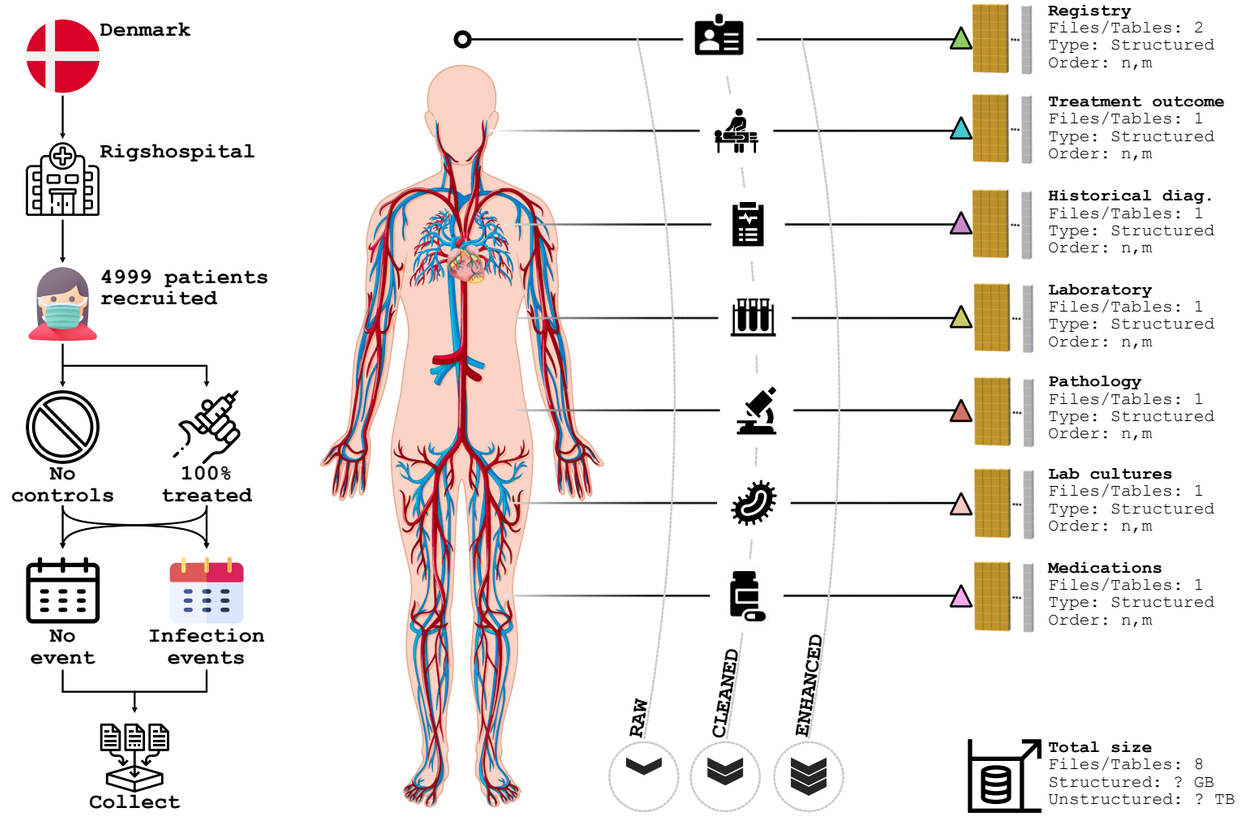
MEDICAL AI DATA SHEET

A principled standard for clinical data communication

CLL-TIM

We aim to both model risk of infection upon CLL treatment and uncover risk factors responsible for low immune function and duration of treatment response upon different treatment regimens.

- No use case disclosed.



COLLECTION:
- Retrospective EHR component
- Prospective trial

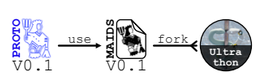
PREPARATION:
- Undisclosed

DISTRIBUTION:
- Protected
- On request
- Collaboration
- Ultrathon 2021

MAINTENANCE:
- Undisclosed

Medical AI DataSheet (MAIDS) v0.1
Website: https://ultrathon.online
Project: github

MAIDS Provenance



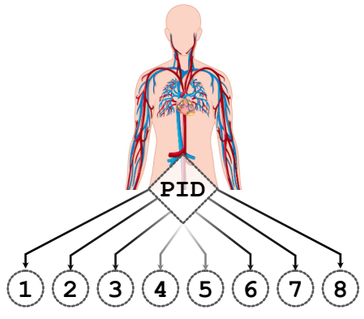
Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Rudi Agius

Description of subsets

Table 1. Available Subsets

SID	Name	Format / Size	Purpose
1	Baseline	CSV/4999	Contains CLL prognostic factor variables taken around time of diagnosis
2	CLL Registry	CSV/4999	Patient information relevant for generaiton of outcomes related to treatment and death
3	Treatment Outcome	CSV/1996	Treatment outcome information that includdes lines of treament and type of treament
4	Diagnosis	CSV/147197	Historical data on previous patient diagnosis - not necessarily related to CLL
5	Laboratory	CSV/4226248	Laboratory tests data - not necessarily related to CLL
6	Pathology	CSV/199337	Historical data on previous patient pathology - not necessarily related to CLL
7	Laboratory Cultures	CSV/84299	Laboratory culture data - necessarily for creating features and outcomes related to infections
8	Medications	CSV/1098273	Historical data on prescribed medications - not neccesarily related to CLL

Subset relationships



Statement of intent

> MOTIVATION

Category 1-of-7 (4 questions)

The questions in this category are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

M1: For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* Infections are the leading cause of mortality in CLL. Risk of Infection is increased upon CLL treatment and currently we have no model that is able to predict risk of infection upon CLL treatment. The dataset created puts together various sources of time-series electronic health records on CLL patients in Denmark. This also includes outcome on death, treatment and infection. Using this data set we aim to both model risk of infection upon CLL treatment and uncover risk factors responsible for low immune function and duration of treatment response upon different treatment regimens. [By: Rudi Agius]

M2: Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)? Dataset was created by the CLL Laboratory at Rigshospitalet, Copenhagen University Hospital using data from the national CLL registry and PERSIMUNE data warehouse. [By: Rudi Agius]

M3: Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.* Novo Nordisk Foundation (grant NNF16OC0019302) and Danish Cancer Society. [By: Rudi Agius]

M4: Any other comments? Answer. [By: Rudi Agius]

> COMPOSITION

Category 2-of-7 (17 questions).

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

C1: What do the instances that comprise the dataset represent (e.g., samples, images, people)? *Are there multiple types of instances (e.g., samples, images, and people), interactions (e.g., nodes and edges), resolutions (e.g., genetic data, single cell expression vs. tissue expression, cell counts, different image technologies, etc.)? Please provide a description.* Dataset represents electronic health record data from various data sources collected on CLL Patients in Denmark. Baseline: Each instance is a single CLL patient and their set of CLL prognostic factor variables taken around time of CLL diagnosis

CLL Registry: Each instance is a single CLL patient and extracted from the National CLL Registry that contains information on several clinically relevant outcome like treatment and death. Treatment Outcome data contains information on line of CLL treatment and type of CLL treatment. Each instance is a line of treatment – hence multiple instances for patients with multiple lines of treatment. For the following each instance is one patient event – hence multiple instances for patients with multiple events – where events are a diagnosis, pathology, laboratory test, prescription, lab culture. Diagnosis: Historical data on previous patient diagnosis - not necessarily those related to CLL. Laboratory: Laboratory tests – not necessarily related to CLL. Pathology: Historical data on previous patient pathology - not necessarily related to CLL. Laboratory Cultures: data for different type of cultures taken for a given patient – thereby holding information on CLL related infections. Medications: Historical data on prescribed medications - not necessarily related to CLL. [By: Rudi Agius]

C2: How many instances are there in total? *Provide an exact integer value for each type mentioned in question C1.* Baseline: 4999 (each instance is a unique patient)

CLL Registry:4999 (each instance is a unique patient)

Treatment Outcome:1996 (No, or multiple instances for each patient possible)

Diagnosis:147197 (No, or multiple instances for each patient possible)

Laboratory:4226248 (No, or multiple instances for each patient possible)

Pathology:199337 (No, or multiple instances for each patient possible)

Laboratory Cultures:84299 (No, or multiple instances for each patient possible)

Medications:1098273 (No, or multiple instances for each patient possible). [By: Rudi Agius]

C3: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this*

Medical AI DataSheet (MAIDS) v0.1
Website: https://ultrathon.online
Project: github
Copyright: CHIP, Rigshospitalet (2021)

MAIDS Provenance
PROTO V0.1 use MAIDS V0.1 fork Ultra thon

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Rudi Agius

representative-ness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., an active decision to cover a more diverse range of instances, because instances were withheld or unavailable). CLL Registry contains data on all CLL patient in Denmark. Remaining datasets have missingness both at random and not at random. [By: Rudi Agius]

C4: What data does each instance consist of? *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.* Each data instance is an instance of a patient record - hence raw data. A single patient is therefore spread across several instances in different datasets. For example, a particular laboratory test, or a single prescription of a medication, will hold an instance each. Features must then be derived from these raw data to generate just a single-row for each patient as is typical in a machine learning dataset. [By: Rudi Agius]

C5: Is there a label, target, or outcome (e.g., mortality) associated with each instance? *If so, please provide a description and indicate its actual presence within the dataset or whether it is represented by a proxy or compounded (e.g., a multi-cause event).* No, as each instance is raw data. Outcomes may be derived from CLL Registry, Treatment Outcome, and Laboratory Cultures datasets. [By: Rudi Agius]

C6: Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.* No information has been redacted. Missing information can be missing at random and not at random. [By: Rudi Agius]

C7: Are relationships between individual instances made explicit (e.g., familial links, or samples derived from the same patient or same exposure)? *If so, please describe how these relationships are made explicit.* Yes. The patient id links different instances in a given dataset to a single patient, and similarly instances from other datasets related to the same patient. [By: Rudi Agius]

C8: Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.* No. [By: Rudi Agius]

C9: Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.* Yes. There are potential duplicates and some of the dates and times may also have errors. Errors may also be present in laboratory values. [By: Rudi Agius]

C10: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, public databases, other datasets and/or private silos)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.* Yes. Datasets is extracted from the National CLL registry and PERSIMUNE data warehouse and continuously updated. [By: Rudi Agius]

C11: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? *If so, please provide a description.* Yes. [By: Rudi Agius]

C12: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.* No. [By: Rudi Agius]

C13: Does the dataset not relate to people (e.g., animals, cell lines, environment)? *A short answer is sufficient. If no relation to people, you may skip the remaining questions in this section.* Direct relation to people. [By: Rudi Agius]

C14: Does the dataset identify any subpopulations (e.g., by age, gender, etc.)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* Age and gender are available and hence any information may be stratified according to these variables. [By: Rudi Agius]

C15: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.* Patient IDs are pseudonymized in all datasets. It is potentially possible to identify individuals indirectly in combination with other data. [By: Rudi Agius]

C16: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.* No. [By: Rudi Agius]

C17: Any other comments? None. [By: Rudi Agius]

> COLLECTION PROCESS (not completed)

Category 3-of-7 (13 questions).

If possible, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior category, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

L1: How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, instrument measurements), reported by subjects/ physicians (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses, scores, etc.)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/ verified? If so, please describe how.* Answer. [By: Surname, name]

L2: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?* Answer. [By: Surname, name]

L3: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? *Please describe.* Answer. [By: Surname, name]

L4: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., salaried, immaterial through prizes / authorship / etc) and how much (e.g., according to competitive scales mandated by [insert body or institution])? Answer. [By: Surname, name]

L5: Over what timeframe was the data collected? *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent data from old biobanked samples, or recent data dump from a 5-year-old registry)? If not, please describe the time frame in which the data associated with the instances was created.* Answer. [By: Surname, name]

L6: Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.* Answer. [By: Surname, name]

L7: Does the dataset not relate to people (e.g., animals, cell lines, environment)? *A short answer is sufficient. If no relation to people, you may skip the remaining questions in this section.* Answer. [By: Surname, name]

L8: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? *Please explain.* Answer. [By: Surname, name]

L9: Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.* Answer. [By: Surname, name]

L10: Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.* Answer. [By: Surname, name]

L11: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).* Answer. [By: Surname, name]

L12: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.* Answer. [By: Surname, name]

L13: Any other comments? Answer. [By: Surname, name]

> PREPROCESSING / CLEANING / LABELING (not completed)

Category 4-of-7 (4 questions).

If possible, dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this category are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks.

P1: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.* Answer. [By: Surname, name]

P2: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g.,

Medical AI DataSheet (MAIDS) v0.1
Website: https://ultrathon.online
Project: github
Copyright: CHIP, Rigshospitalet (2021)

MAIDS Provenance
PROTO V0.1 use MAIDS V0.1 fork Ultra thon

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Rudi Agius

to support unanticipated future uses)? *If so, is it available and needs to be done to gain access? If open without restriction then please describe a means to access this "raw" data.* Answer. [By: Surname, name]

P3: Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point and describe with enough detail so that others might reproduce it. If a custom script was used will you include it within the MAIDS repository or otherwise make it available.* Answer. [By: Surname, name]

P4: Any other comments? Answer. [By: Surname, name]

> USES (not completed)

Category 5-of-7 (6 questions).

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harm.

U1: Has the dataset been used for any tasks already? *If so, please provide a description. A detailed response will help others determine the value of this dataset by example.* Answer. [By: Surname, name]

U2: Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point. Will you compile such a list and make it available in the MAIDS repository.* Answer. [By: Surname, name]

U3: What (other) tasks could the dataset be used for? *Please provide as much inspiration as you can. Distinguish between tasks the dataset is ideal for versus those tasks where the dataset is not entirely suited. Describe why the dataset might not be suitable.* Answer. [By: Surname, name]

U4: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?* Answer. [By: Surname, name]

U5: Are there tasks for which the dataset should not be used? *If so, please provide a description.* Answer. [By: Surname, name]

U6: Any other comments? Answer. [By: Surname, name]

> DISTRIBUTION (not completed)

Category 6-of-7 (7 questions).

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

D1: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description. If not, then disregard the rest of the questions.* Answer. [By: Surname, name]

D2: How will the dataset be distributed (e.g., tarball on website, API, GitHub)? *Does the dataset have a digital object identifier (DOI).* Answer. [By: Surname, name]

D3: When will the dataset be distributed? *A cautious response is more useful than an optimistic one.* Answer. [By: Surname, name]

D4: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.* Answer. [By: Surname, name]

D5: Have any third-parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.* Answer. [By: Surname, name]

D6: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.* Answer. [By: Surname, name]

D7: Any other comments? Answer. [By: Surname, name]

> MAINTENANCE (not completed)

Category 7-of-7 (8 questions).

As with the previous category, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

T1: Who is supporting/hosting/maintaining the dataset? *Please be as thorough as possible.*

Answer. [By: Surname, name]

T2: How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Answer. [By: Surname, name]

T3: Is there an erratum? *If so, please provide a link or other access point.* Answer. [By: Surname, name]

T4: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub).* Answer. [By: Surname, name]

T5: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.* Answer. [By: Surname, name]

T6: Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.* Answer. [By: Surname, name]

T7: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.* Answer. [By: Surname, name]

T8: Any other comments? Answer. [By: Surname, name]

Medical AI DataSheet (MAIDS) v0.1
Website: https://ultrathon.online
Project: github
Copyright: CHIP, Rigshospitalet (2021)

MAIDS Provenance
PROTO V0.1 → use → MAIDS V0.1 ← fork ← Ultra thon

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Rudi Agius