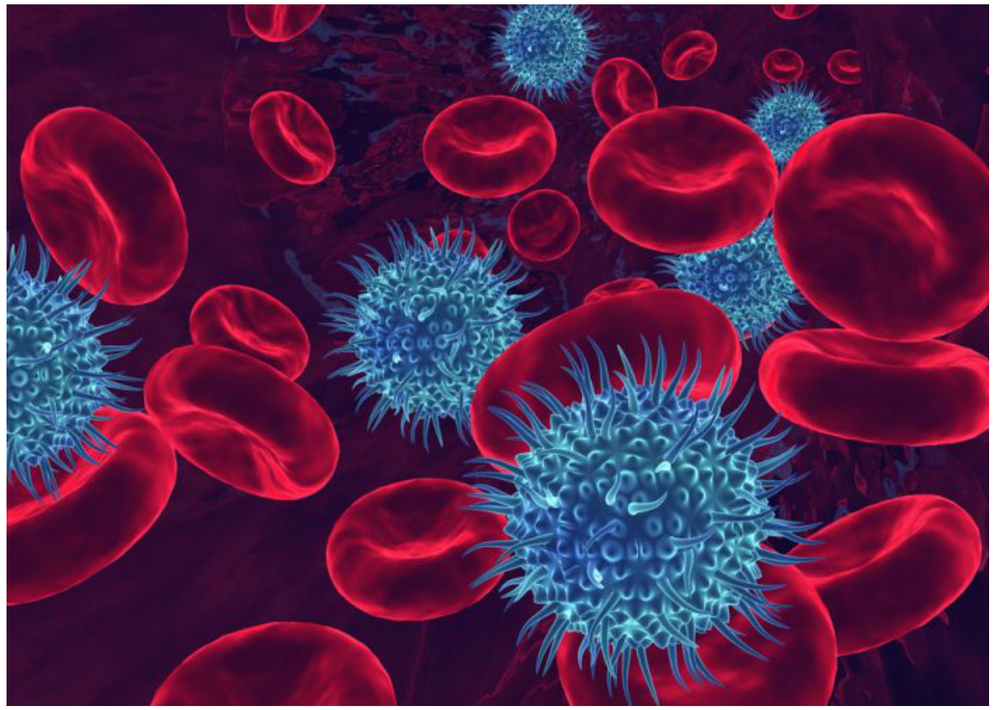
Copenhagen Ultrathon on Precision Medicine

Karen Leth Nielsen
Frederik Boëtius Hertz
Niels Frimodt-Møller
Steen Rasmussen
Ruth Frikke-Schmidt
Jesper Qvist Thomassen

Blood culturing data associated to clinical metadata

Challenge ID: U21-01



Main research question:
Bloodstream infections (BSI) are a major cause of morbidity and mortality (mortality rates up to 30%) with increasing incidence. At Rigshospitalet, Copenhagen, 40,000 blood culturing flasks are sampled every year when suspecting BSI. Of these, only 10% are true positives and 10% contain contaminants, often skin flora, i.e. prediction of BSI is poor.

While samples are cultured for identification of pathogens, broad empirical antibiotic treatment (BEAT) is initiated, which increases risk of antimicrobial resistance and subsequent acquisition of multidrug resistant pathogens. When culturing is finalized days after sampling BEAT is changed to a narrow treatment to specifically target the identified pathogen. More precise prediction of BSI-positive patients including identification of the infecting pathogen and application of more precise treatment few hours after suspicion of BSI would be of great advantage for all patients with increased survival, less development of resistance, fewer side effects and be economically beneficial for the hospital.

The aim is to employ machine learning to design an AI-based clinical application which can predict bloodstream infections from clinical parameters which are measured right upon admission. The present dataset contains:

- Positive and negative blood culturing results of vira, bacteria and fungi
- any previous microbial culturing results up to six months prior to BSI suspicion
- Patient age and Sex
- previous admissions (departments, dates)
- diagnoses
- biochemical blood-analyses (electrolytes, glucose, creatinine, infection parameters [leukocytes, CRP, procalcitonin], eGFR and liver enzymes).

Secondary research question(s):
Is it possible to predict not only BSI but also the infecting pathogen already a few hours after admission without the culturing result?

KEY

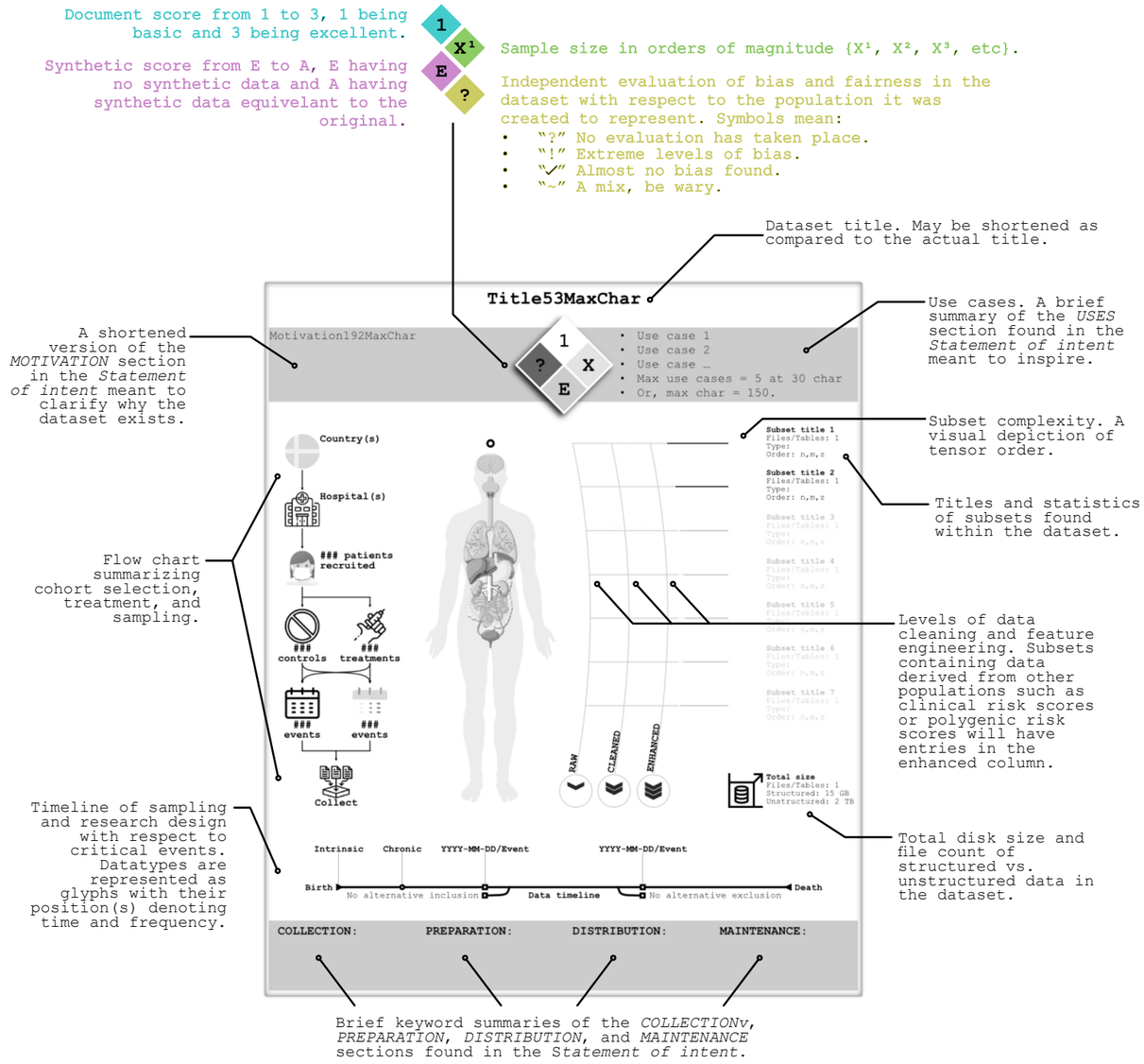
Document score from 1 to 3, 1 being basic and 3 being excellent.

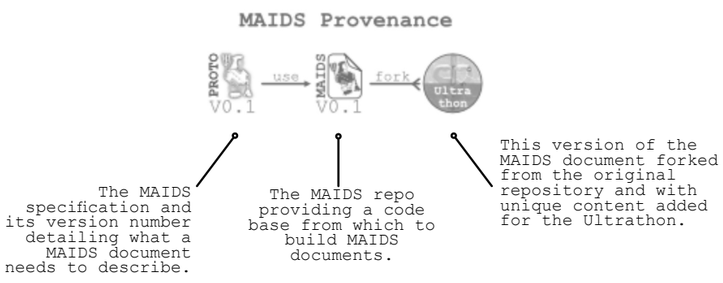
Synthetic score from E to A, E having no synthetic data and A having synthetic data equivelant to the original.

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.





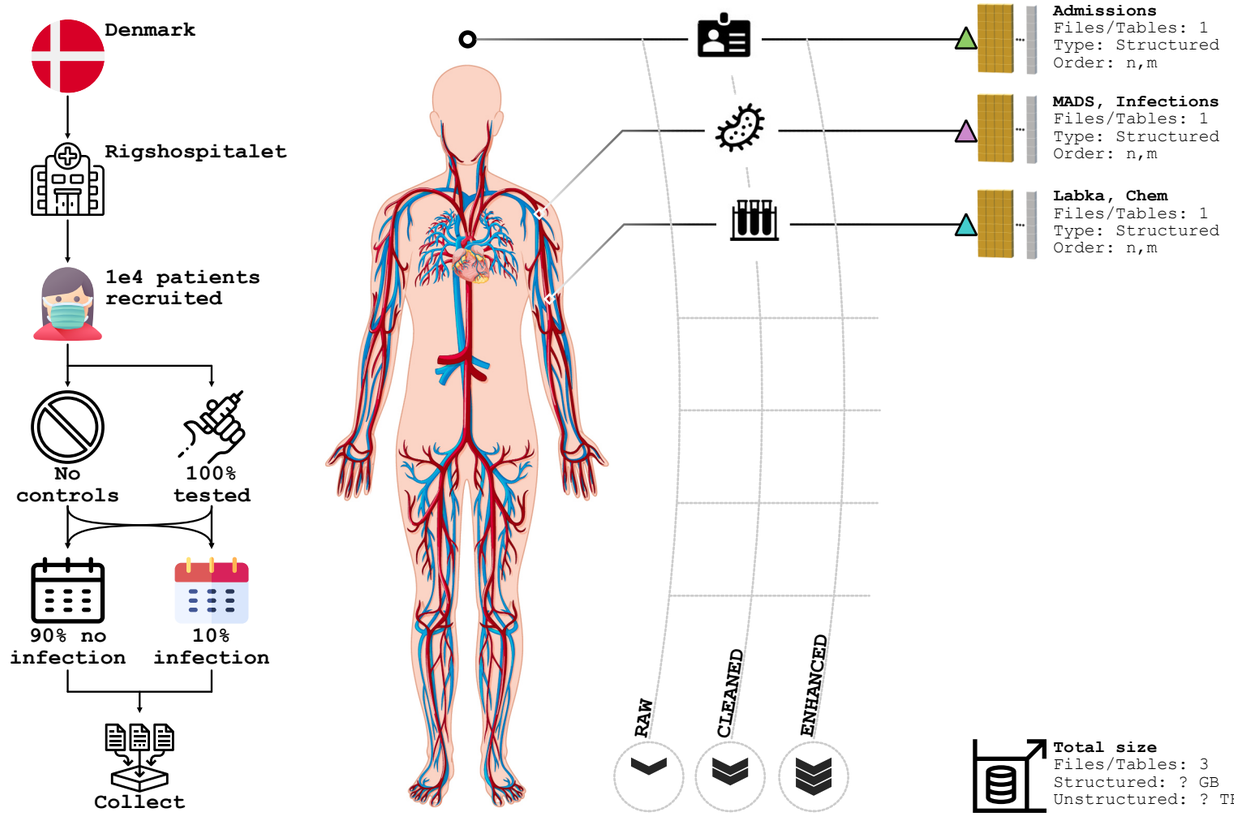
MEDICAL AI DATA SHEET

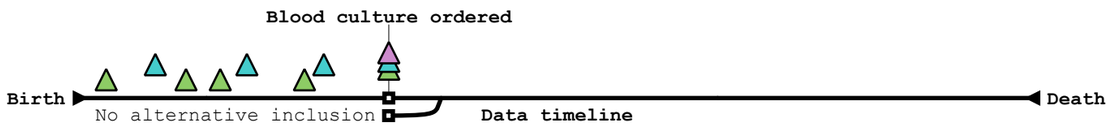
A principled standard for clinical data communication

BLOOD CULTURING DATA ASSOCIATED TO CLINICAL METADATA

The present dataset was created for machine learning in order to improve the prediction of blood-stream infections.

- No use cases disclosed.





COLLECTION:
- Directly observed.
- Retrospective, 10-year lookup.

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

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

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Karen Leth Nielsen

Description of subsets

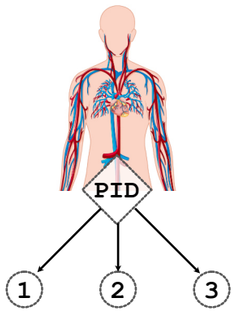
Table 1. Available Subsets

SID	Name	Modality / Format / Size	Purpose
1	**MADS**	Structured / CSV / 1e4	The positive/negative blood culturing results are essentieal for the prediction of blood stream infections.
2	**Admissions**	Structured / CSV / 1e4	This data informs us of when the patient was admitted and is relevant for filtering the clinical blood values to only iclude the ones relavant for the BSI. Additionally, previous admissions can be predictive for BSI development.
3	**Labka**	Structured / CSV / 1e4	The clinical blood values are important for the prediction of BSI.

Table 2. Definitions and keywords

KID	Keyword	Definition
1	BSI	Bloodstream infection.

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.* The present dataset was created for machine learning in order to improve the prediction of bloodstream infections. The dataset combines clinical blood parameters, age, sex as well as previous admissions to blood culturing results (negative and positive) and resistance profile of the infecting pathogen over a period of more than 10 years. [By: Karen Leth Nielsen]

M2: Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?. Dept. of Clinical Microbiology, Rigshospitalet created the dataset In collaboration with dept. of clinical Biochemistry, Rigshospitalet. [By: Karen Leth Nielsen]

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.* Undislosed. [By: Karen Leth Nielsen]

M4: Any other comments? None. [By: Karen Leth Nielsen; Frederik.boetius.hertz@regionh.dk]

> 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.* The instances comprise positive and negative findings of all blood culturing flasks taken in the period 2010-2020 at Rigshospitalet. For these patients admission history and current admission was collected together with biochemical blood sampling data. [By: Karen Leth Nielsen]

C2: How many instances are there in total? *Provide an exact integer value for each type mentioned in question C1.* Answer. [By: Surname, name]

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 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).* The dataset has not been filtered and contains all possible instances. [By: Karen Leth Nielsen]

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 instance consists of a blood culturing flask which is either positive or negative for bacteriological findings. These data are combined with blood values from the biochemical database of the same individual as well as patient details and admission data. [By: Karen Leth Nielsen]

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).* Answer. [By: Surname, name]

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.* Answer. [By: Surname, name]

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.* The dataset contains multiple samples from the same patients. The pseudonym makes this explicit. [By: Karen Leth Nielsen]

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.* Answer. [By: Surname, name]

C9: Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.* Answer. [By: Surname,

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: Karen Leth Nielsen

name]

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.* The data is self-contained. [By: Karen Leth Nielsen]

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.* Answer. [By: Surname, name]

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: Karen Leth Nielsen]

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.* Answer. [By: Surname, name]

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.* Answer. [By: Surname, name]

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.* It is not possible to identify individuals with this dataset alone. [By: Karen Leth Nielsen]

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.* Yes: The dataset contains health data including previous admissions and blood values indicating general health. It is not possible to identify individuals behind from the dataset alone. [By: Karen Leth Nielsen]

C17: Any other comments? No. [By: Karen Leth Nielsen]

> COLLECTION PROCESS

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.* Microbial and biochemical blood analyses, hence, directly measured or observed using validated instruments and methods. [By: Karen Leth Nielsen]

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?* The biochemical blood values were analyzed using standard hospital assays. Data was stored in the hospital laboratory information system. The relevant data were extracted using Oracle SQL based on personal identification number, analysis codes and sampling dates. The microbiological data was analysed with standard hospital procedures and stored in MADS. Extracted with SAS. Admission and other patient data were extracted with SAS. [By: Jesper Qvist Thomassen, Steen Rasmussen]

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.* This dataset is the full dataset. [By: Karen Leth Nielsen]

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])? The data was collected by the data owners. [By: Karen Leth Nielsen]

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. Data dump from a 11-year old registry, created March 2021. [By: Karen Leth Nielsen]

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.* No. [By: Karen Leth Nielsen]

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.* Direct relation to people. [By: Karen Leth Nielsen]

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.* The data was collected from the hospital databases and regional databases. [By: Karen Leth Nielsen]

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.* No, the data was collected retrospectively. [By: Karen Leth Nielsen]

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.* No. [By: Karen Leth Nielsen]

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).* NA. [By: Karen Leth Nielsen]

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.* No. [By: Karen Leth Nielsen]

L13: Any other comments? No. [By: Karen Leth Nielsen]

> PREPROCESSING / CLEANING / LABELING

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.* The biochemical blood value data was filtered such that only valid test answers were included. Hence, instances where the specific analysis failed for whatever reason, the data was filtered out. [By: Jesper Qvist Thomassen]

P2: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., 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.* The "raw" data is stored in the hospitals laboratory information system. [By: Jesper Qvist Thomassen]

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.* Yes. For biochemical blood values: Oracle SQL. For all other data: SAS, SAS/CONNECT, SAS/ACCESS. [By: Jesper Qvist Thomassen, Steen Rasmussen]

P4: Any other comments? No. [By: Jesper Qvist Thomassen]

> USES

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.* No. [By: Karen Leth Nielsen]

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]

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 → Ultrathon

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Karen Leth Nielsen

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.* The dataset is ideal for analyzing explanatory factors for prediction of BSI. In addition, the biochemical blood values might be able to predict infection and the infecting pathogen alone, which can be analyzed from the current dataset. Other applications include analyzing outcome of the patients in relation to empiric antibiotic treatment and the timing of this. Prediction of explanatory factors for death. Antibiotic treatment and eGFR. What is the outcome of patients with a negative blood culturing result have? Can we identify parameters which clear them from BSI suspicion so the empiric treatment can be stopped?

Not entirely suitable: Effect on inflammatory markers in correlation to antibiotic treatment. Not entirely suited because many patients are prescribed several antibiotics. Can only be done for a subset of the patients. [By: Karen Leth Nielsen]

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?* No. [By: Karen Leth Nielsen]

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? No. [By: Karen Leth Nielsen]

> DISTRIBUTION

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.* No. [By: Karen Leth Nielsen]

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 DATA SHEET
A principled standard for clinical data communication

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 Ultrachon

Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Karen Leth Nielsen