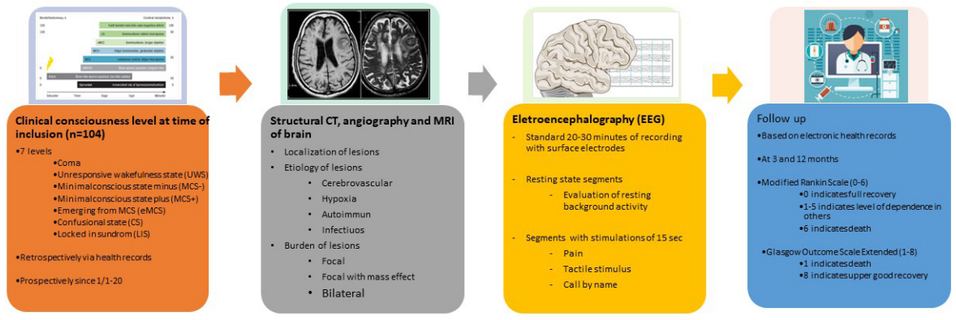
Copenhagen Ultrathon on Precision Medicine

Daniel Kondziella
Moshgan Amiri

Consciousness in neurocritical care cohort study

Challenge ID: U21-03



Main research question: After acute brain injury, recovery of consciousness is the single most important factor for clinical outcome. Consciousness levels in clinically unresponsive patients are misdiagnosed in 15-20%, i.e. consciousness may be under- or overestimated, which has negative effects on clinical decision making in the intensive care unit (ICU). Thus, underestimation may lead to premature withdrawal of life-supporting therapy, whereas overestimation may lead to prolongation of futile treatment. Given that 7 out of 10 deaths in the ICU occur because of treatment withdrawal, accurate estimation of consciousness is crucial to save patients from poor medical decision making.

The main objective of this ongoing multimodal study is to facilitate individualized assessment of unresponsive patients with disorders of consciousness in the ICU for signs of preserved consciousness.

Since 2015, we have systematically and rigorously registered data regarding clinical neurological exam (including detailed scoring of consciousness levels), laboratory investigations including structural brains scans (computed tomography and magnetic resonance imaging), electroencephalography (EEG), cerebrospinal fluid samples and other relevant exams during ICU admission. Furthermore, to obtain information on recovery after acute brain injury, we registered follow-up data at 3- and 12 months according to modified Rankin scale (mRS) and Glasgow outcome scale extended (GOS-E), or alternatively, the cause of death during ICU admission.

As of January 2021, we have included 104 patients. Follow-up data at 3- and 12 months are available from 78 and 70 patients, respectively.

We hypothesize that AI can uncover important signatures of consciousness levels and cognitive outcome in this dataset that escapes traditional interpretations by physicians.

Does our clinical database contain information about consciousness levels and clinical outcomes that we clinicians do not recognize but that AI can identify?

Secondary research question(s):

1. Can machine learning algorithms recognize common signatures indicating good or poor outcome during ICU admission, in an inhomogeneous group of patients with acute brain injury, based on clinical and laboratory features which are missed by clinical routine evaluation?

2. Can machine learning algorithms assist clinicians in a more accurate decision-making regarding end-of-life decisions in the ICU?

3. Can machine learning algorithms predict outcome in unresponsive patients with acute brain injury at ICU discharge and 3 and 12 months later?

4. Is there a subset of unresponsive patients with acute brain injury for whom AI evaluation is particularly helpful?

5. Will such algorithms be a convenient and cost-efficient for the detection of preserved consciousness, prognostication and treatment of patients with acute brain injury?

KEY

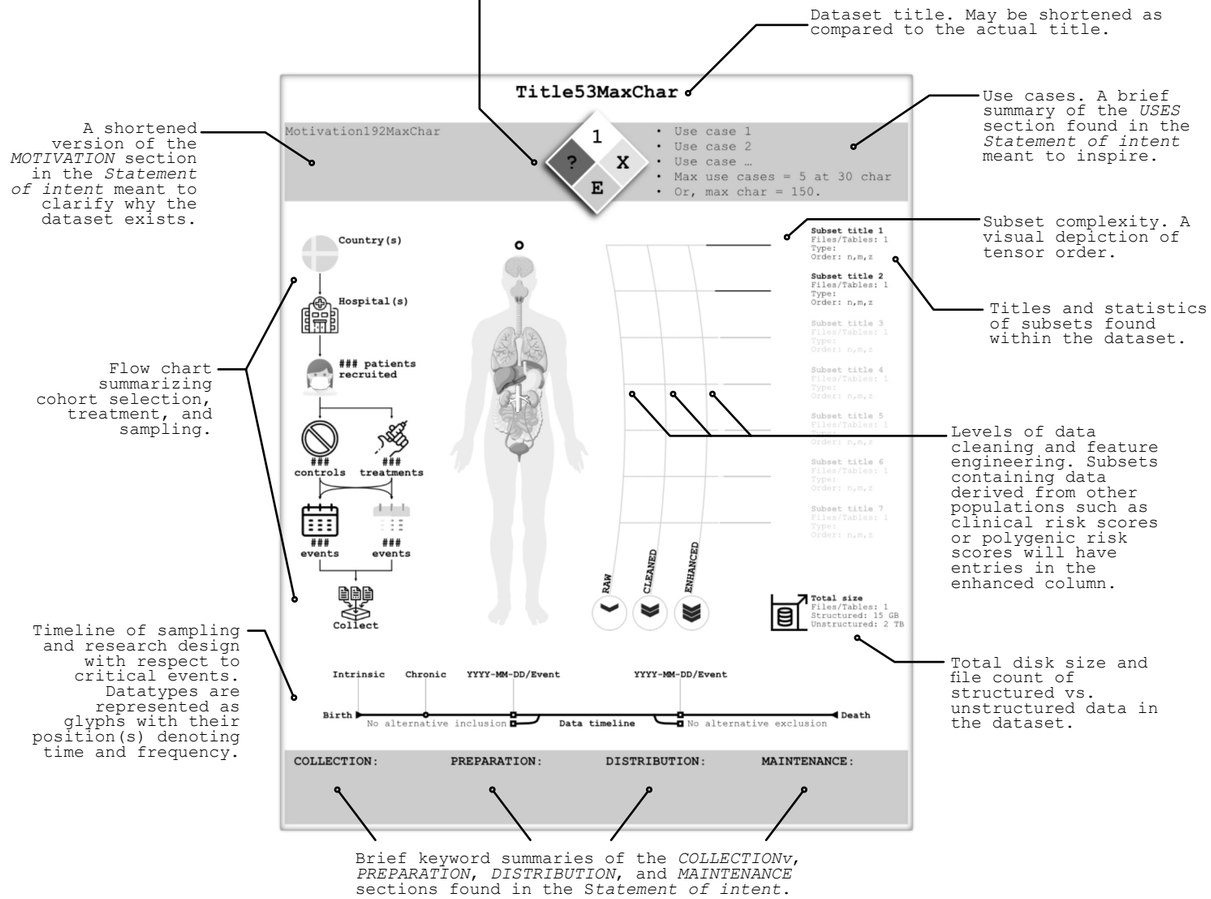
Document score from 1 to 3, 1 being basic and 3 being excellent.

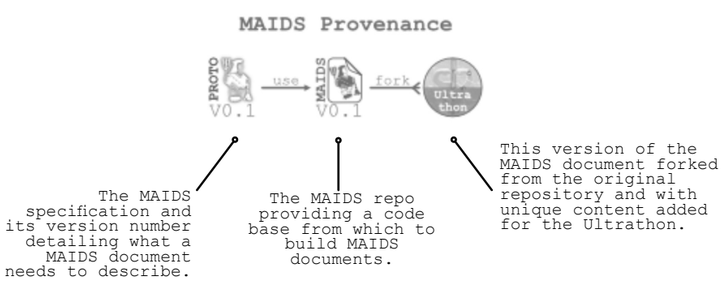
Synthetic score from E to A, E having no synthetic data and A having synthetic data equivelant to the original.

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.





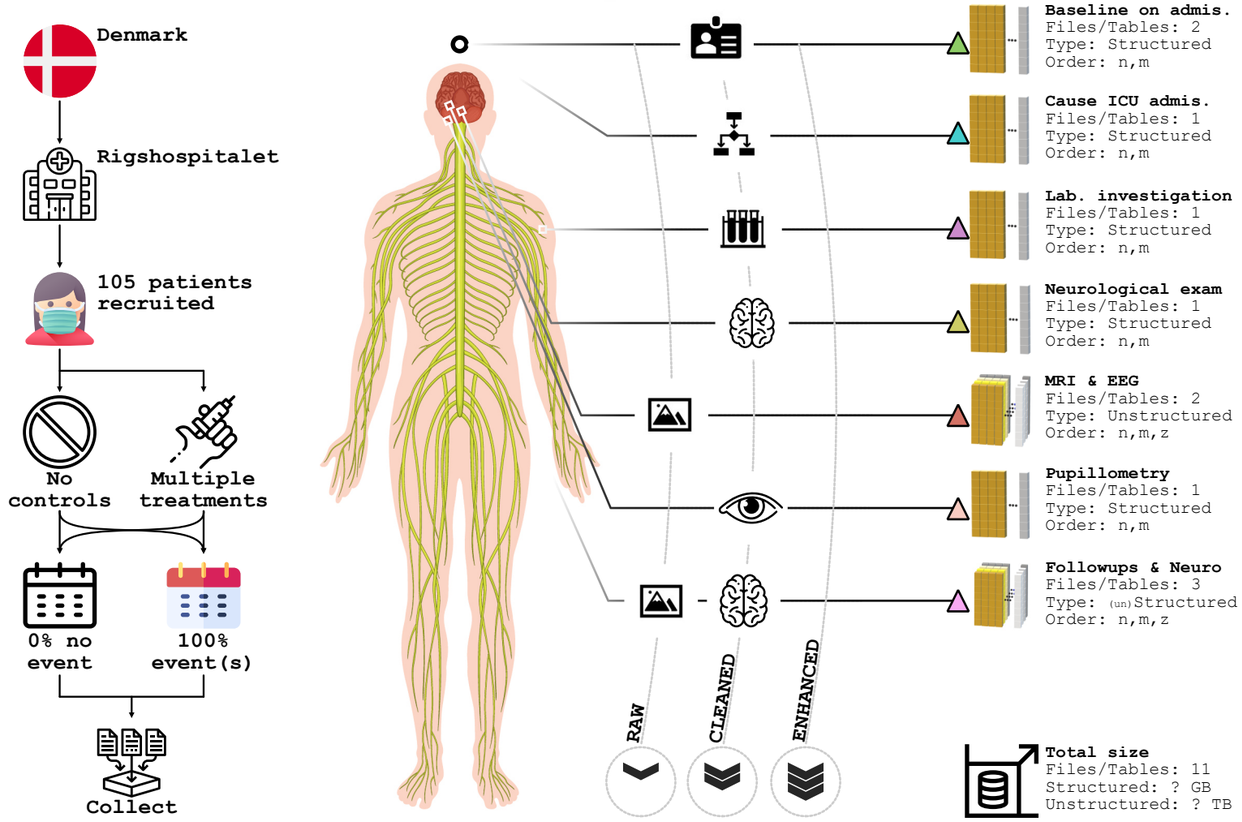
MEDICAL AI DATA SHEET

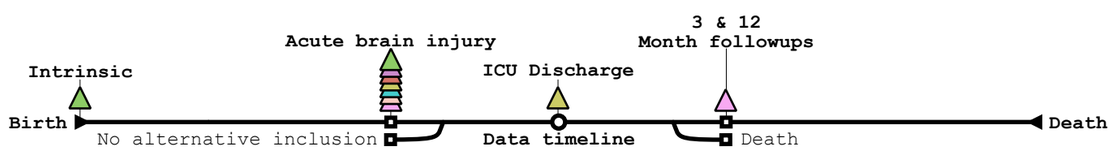
A principled standard for clinical data communication

Consciousness in neurocritical care cohort study

There is a gap between the research on chronic patients and research based on unresponsive patients suffering from acute brain injury in the ICU.

- No use cases disclosed.





COLLECTION:
- PROSPECTIVE
- OBSERVED
- FOLLOWUP

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: Daniel Kondziella

Description of subsets

Table 1. Available Subsets

SID	Name	Format / Size	Purpose
1	Study ID	CSV / 104	Project ID and date of enrolment
2	Cause of ICU admission	CSV / 104	Describing primary cause of admission, date of admission and date of injury
3	Baseline data at ICU admission	CSV / 104	Demographics (age, sex, previous medical history, previous level of daily function), overall clinical condition at admission (temperature, level of creatinine level of movement, early hypoxia or hypotension)
4	Laboratory investigations	CSV / 104	Describing results of brain scans, X-ray, blood tests, electroencephalography and cerebrospinal fluid tests done during the admission
5	Neurological exam at inclusion	CSV / 104	Clinical neurological exams with testing of brainstem reflexes, motor output, consioussness level from a clinical point of view, reaction to stimulies like pain, voice etc.
6	MRI	CSV / 104	Describing structural MRI of brain with sites of lesion, form of lesions (hemmorhage, ischemia, diffus contusion etc.), which sequences were done and final conclusion of the scan described by a radiologist
7	EEG	CSV / 104	Describing brain derived electrical signals measured by surface electrodes and response to different stimuli.
8	Pupillometry	CSV / 104	Descriping pupil characteristics (size, difference between the two pupils etc) and response to light
9	Neurological exam at ICU discharge	CSV / 104	Clinical neurological features and consciousnes level at time of discharge. Cause and date of death if patients die during ICU admission. Final diagnosis.
10	3-month follow up	CSV / 104	Description of daily function regained and level of indepency after admission at 3 months described by 3 different scales (CPC, GOS-E and mRS)
11	12-month follow up	CSV / 104	Description of daily function regained and level of indepency after admission at 12 months described by 3 different scales (CPC, GOS-E and mRS)

Subset relationships

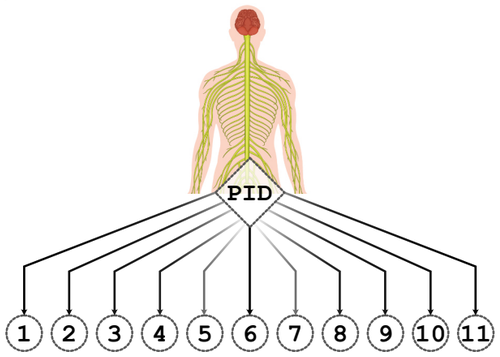


Table 2. Definitions & Keywords

KID	Keyword	Definition
1	ICU	Intensive Care Unit
2	mRS	modified Rankin Scale
3	ICH	Interacerebral Hemorrhage
4	SAH	Subarachnoid Hemorrhage
5	GCS	Glasgow Coma Scale
6	FOUR	Full Outline of UnResponsiveness
7	CPC	Cerebral Performance Cathegory
8	GOS-E	Glasgow Outcome Scale Extended
9	WFNS	World Federation of Neurosurgical Surgeons grading
1v0	CTA	CT angiography
11	ASPECT	Alberta Stroke Program Early CT score
12	DSA	Digital Subtraction Angiography
13	TICI	Thrombolysis In Cerebral Infarction scale
14	IVH	Intraventricular Hemorrhage
15	SDH	Subdural Hemorrhage
16	EDH	Epidural Hemorrhage
17	EEG	Electroencephalography
18	STESS	Status Epilepticus Severity Score
19	EMSE	Epidemiology-based Mortality score in Status Epilepticus
20	ECG	Electrocardiography
21	RASS	Richmond Agitation-Sedation Scale
22	EEG	Electroencephalography
23	MRI	Magnetic resonans Imaging
24	NPi	Neurological pupil index (normal value > 3, value <3 indicates abnormal pupil reaction)
25	ICD-10	Classification of Diseases version 10

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: Daniel Kondziella

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.* Research on consciousness is mainly based on patients suffering from chronic brain injury, while data regarding unresponsive patients with acute brain injury are sparse. As mentioned in the research question, most deaths in an ICU population occur because of withdrawal of life-sustaining therapy. Reducing the risk for erroneous clinical prognostication is therefore crucial. There is a gap between the research on chronic patients and research based on unresponsive patients suffering from acute brain injury in the ICU.

We established our database to be able to fill this gap.

Our dataset was hence created to obtain a representative prospective database with systematic registration of clinical, laboratory and imaging data of unresponsive patients suffering from acute brain injury. The main purpose of the database is to identify important information predicting level of consciousness and outcome (acute and long-term) in these patients, which will help to optimize clinical decision-making. [By: Moshgan Amiri]

M2: Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?. Daniel Kondziella, MD PhD FEBN; Principal investigator of and project developer of CONNECT-ME; Neurocentret, Rigshospitalet. [By: Moshgan Amiri]

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.* Copenhagen University; Offerfonden; Jens Juhl fonden; Rigshospitalets forskningspuljer. [By: Moshgan Amiri]

M4: Any other comments? All our data are derived from electronic medical records, thus if additional details are needed in the process, we will be able to retrieve these from validated records. [By: Moshgan Amiri]

> COMPOSITION (not completed)

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.* Answer. [By: Surname, name]

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

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.* Answer. [By: Surname, name]

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.* Answer. [By: Surname, name]

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

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.* Answer. [By: Surname, name]

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.* Answer. [By: Surname, name]

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.* Answer. [By: Surname, name]

C17: Any other comments? Answer. [By: Surname, name]

> 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*

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: Daniel Kondziella

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., 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 ← Ultrathon

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