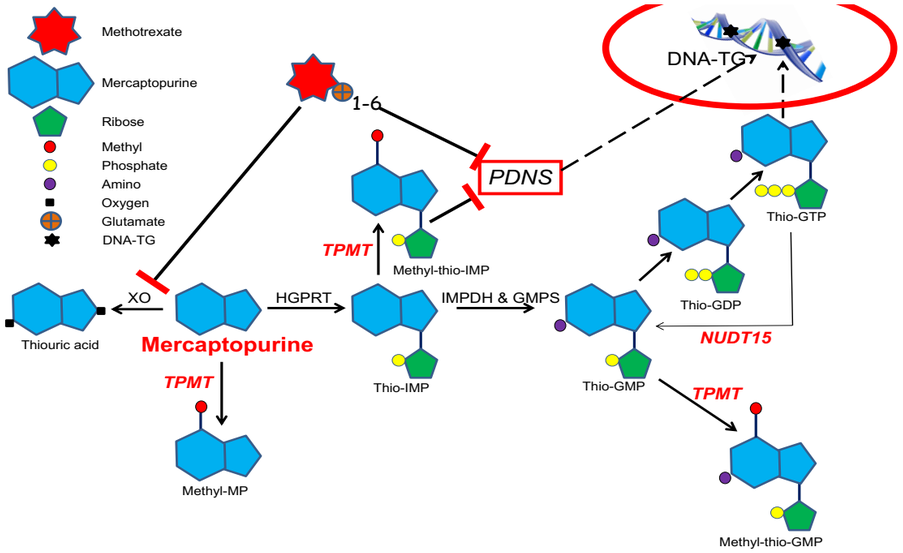
Copenhagen Ultrathon on Precision Medicine

Rikke Linnemann Nielsen
Kjeld Schmiegelow
Kathrine Grell

Thiopurine/methotrexate maintenance therapy of acute lymphoblastic leukemia

Challenge ID: U21-05



Main research question: 1-2 years of thiopurine (6MP)/methotrexate (MTX) maintenance therapy (MT) is one of the most important treatment phases of childhood acute lymphoblastic leukemia without which 40% of all patients will develop leukemic relapse. The pharmacology of MTX and 6MP is complex (see figure), and we are the world-wide leading group in mapping the metabolite landscape in large clinical trials.

Traditionally, MT have been adjusted by blood counts to obtain a preset degree of myelotoxicity, but this is confounded by wide natural variations in blood counts. Thus, patients are currently seen at 1-2 weeks intervals during MT to titrate therapy to the right dosage.

Furthermore, methylated thiopurine metabolites can cause liver cell damage with a rise in aminotransferases (although rarely liver dysfunction; Nygaard, Clin Pharm Ther 2004).

The antileukemic effect of MT is mediated by DNA-incorporated thioguanine nucleotides (DNA-TG) with on average 1:6,000 nucleotides being 6TG substituted. DNA-TG levels varies 10-fold between patients, and DNA-TG incorporation can be enhanced by other thiopurine and methotrexate metabolites. We have recently demonstrated that the risk of relapse is primarily determined by DNA-TG levels during MT (Nielsen, Lancet Oncol 2017; Toksvang, Leukemia 2021 (In press)), and that germline DNA variants determines propensity for DNA-TG incorporation (Tulstrup, Leukemia 2018). However, it frequently takes up to a year before clinicians reach the optimal dose for the individual patient, and drug dosing is determined both by the individual patient's tolerance to the therapy and the treating physician's willingness to adjust the treatment vigorously. Current drug dosing guidelines are based on blood counts (=degree of myelosuppression) but they do not meet the needs and patients are treated very diversely with >10-fold differences in drug doses and 40% do not reach their target degree myelosuppression, which may increase the risk of leukemic relapse. Thus, there is a profound need for AI-driven guidelines that takes into account drug metabolism (of MTX and 6MP), empiric data on drug doses and blood counts across hundreds of patients, and the individual patients treatment trajectory.

Objectives of the present study: To reach a deeper understanding of interactions of these parameters as well as common host genome variants in order to facilitate drug dosing with two aims: (i) more rapidly reach the individual patient's optimal target dose AND with fewer outpatient visits, and (ii) ultimately reduce the risk of relapse due to insufficient drug exposure.

Secondary research question(s): Single nucleotide polymorphisms have been linked to MTX and 6MP metabolism (Moriyama, Nat Genet 2016; Tulstrup, Leukemia 2018; Tulstrup, Blood 2020).

Secondary aim: To develop of a polygenic risk score that reliably can predict metabolism of MTX and 6MP in the individual patients.

Maintenance therapy risk factors predictive of relapse

NOPHO ALL92: 532 non-DS ALL patients (1.0-14.9 years); ~28,000 data set

Those in **red** are included in the final multivariate model

	Beta	p
Male gender	**2.00**	**0.004**[2]
Age at Dx	1.1	0.04
WBC at Dx*	**1.04**	**0.048**[2]
TPMT-phenotype**	**2.7**	**0.03**[1]
6MP dose	0.99	0.15
MTX dose	1.09	0.04
E-TGN	1.0	0.37
E-MTX	0.96	0.45
Leukocyte count	0.76	0.43
Neutrophil count	1.7	**0.0007**[2]
Lymphocyte count	1.8	0.19
Thrombocyte count	1.0	0.53
Hemoglobin	1.0	0.54
Aminotransferases (N=385)	1.16	0.56[3]

Base line parameters

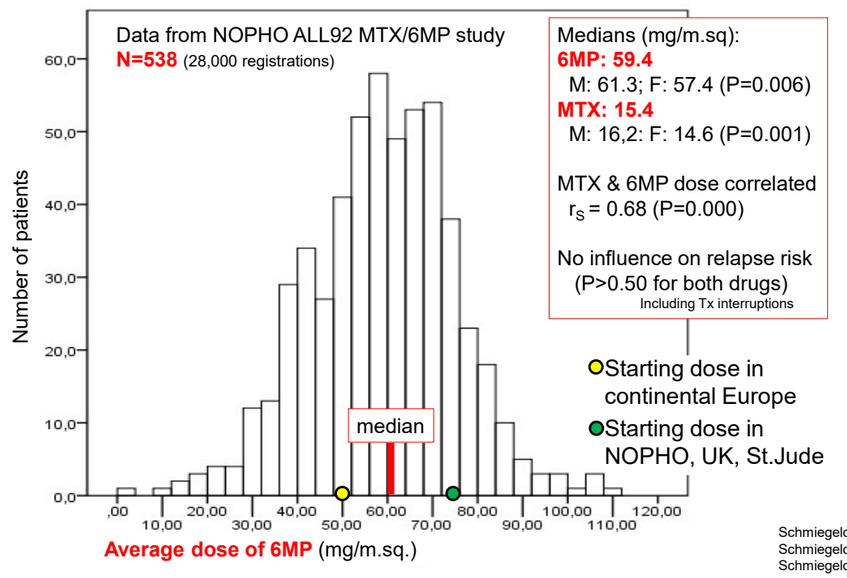
Maintenance therapy parameters

ANC (0.75-1.5) is the target in ALLTogether study
BUT: [4]
Normal ANC varies among patients
Fluctuates and is influenced by
age, ethnicity, time of day, season

*Per 10x10⁹/L
**Sign in final multivariate model

[1] Schmiegelow, Ped Blood Cancer 2016
[1] Schmiegelow, JPHO 2014
[2] Schmiegelow, Leukemia 2009
[3] Ebbesen, Ped Blood Cancer 2017
[4] Nielsen, Cancer Chemother Pharm 2016

Average *prescribed* 6MP doses during maintenance (protocol dose: 75 mg/m.sq.)



Schmiegelow, JCO 2003
Schmiegelow, JPHO 2014
Schmiegelow, PBC 2016

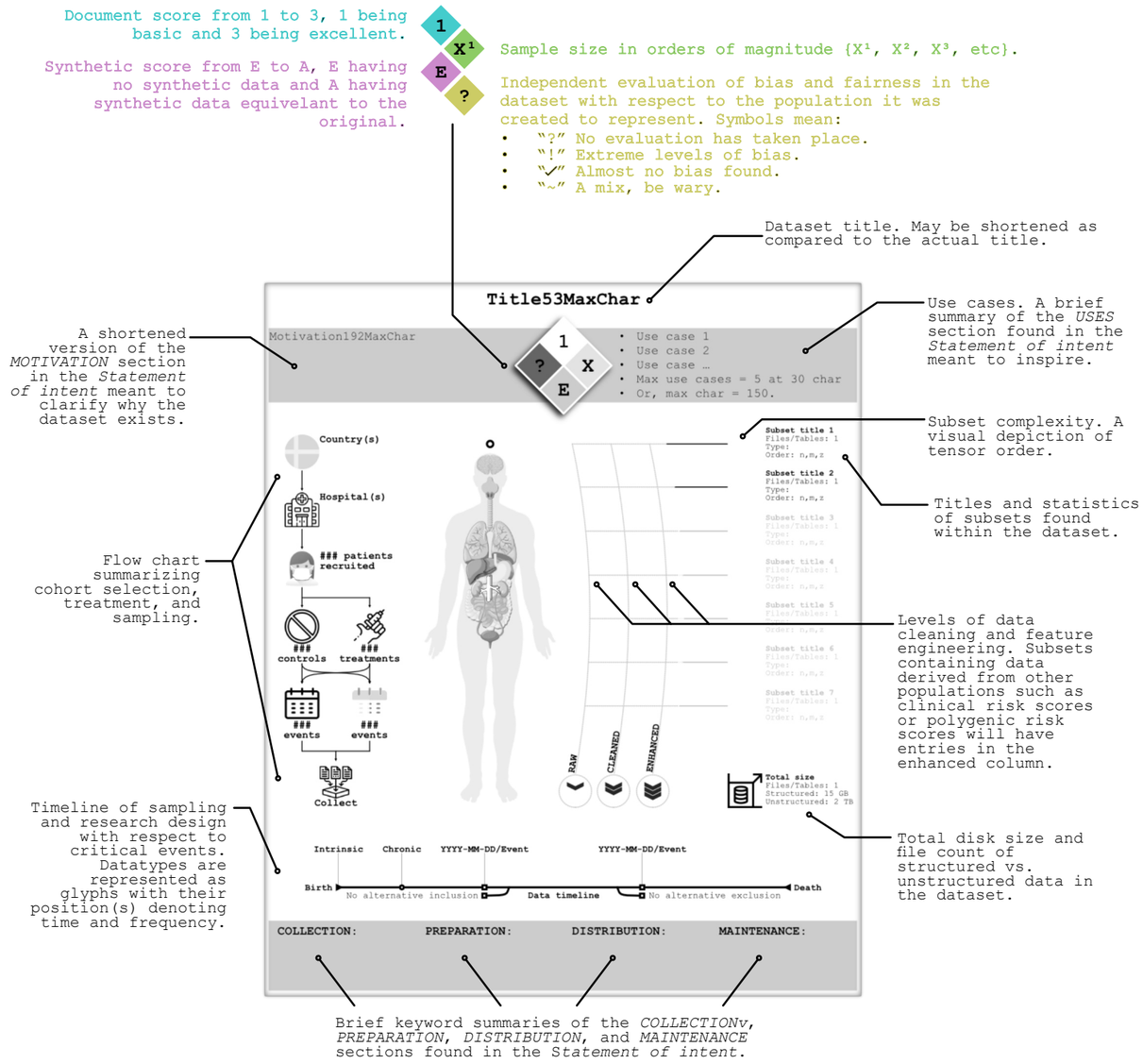
KEY

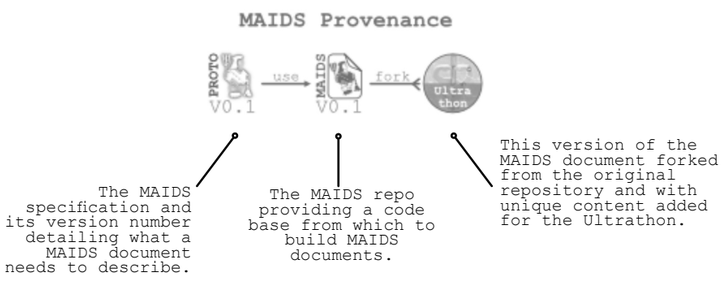
Document score from 1 to 3, 1 being basic and 3 being excellent.

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.





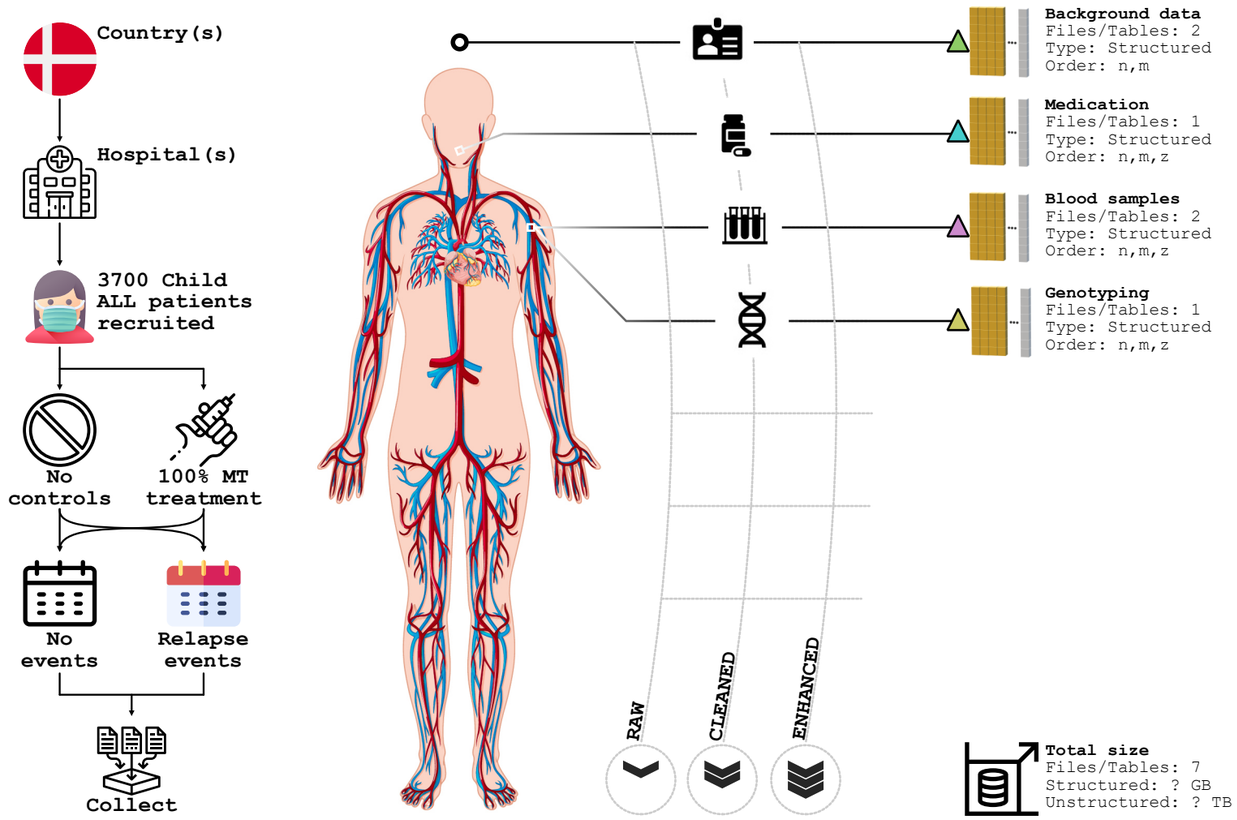
MEDICAL AI DATA SHEET

A principled standard for clinical data communication

MAINTENANCE THERAPY OF ACUTE LYMPHOBLASTIC LEUKEMIA

To provide a deeper understanding of pharmacogenetics/-kinetics/-dynamics of MT and increase cure rates for childhood leukemia.

- No use case disclosed.



COLLECTION:
- Prospective
- Direct observ.
- Genotyping
- Longitudinal

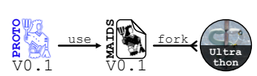
PREPARATION:
- Undisclosed

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: Rikke Linnemann Nielsen

Description of subsets

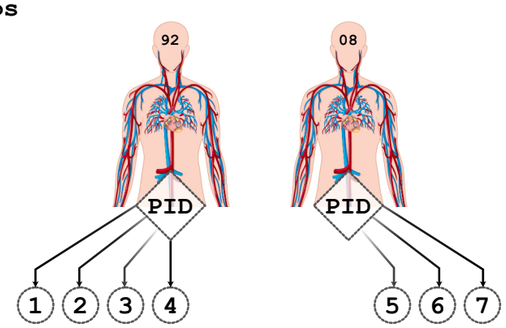
Table 1. Available Subsets

SID	Name	Modality / Format / Size	Purpose
1	ptdata92	sav / 538	NOPHO ALL-92, patient cohort with background information. Follow-up until 20.04.04
2	prdata92	sav / 9209	NOPHO ALL-92, blood samples, one row per sample per patient to determine EMTX and E6TGN levels
3	meddata92	sav / 28582	NOPHO ALL-92, one line per patient reporting medicine, medicine dose or blood sample to determine leukocyte count
4	snitdata92	sav / 538	NOPHO ALL-92, Contains a line for each patient with various averages and patient-specific information. The majority of the averages are taken from the course file and how they are determined therefore appears from the description of the course file.
5	ptdata08	sav / 3162	NOPHO ALL-2008, patient cohort with background information.
6	prdata08	xlxs / 43896	NOPHO ALL-08, blood samples, one row per sample per patient to determine metabolite levels incl. Methylated metabolites
7	Genotype data08	Binary plink / 3	NOPHO ALL-2008. Genome-wide SNP profiling. 2146021 variants and 1829 people pass filters and QC.

Table 2. Definitions & Keywords

KID	Keyword	Definition	Links
1	EMTX	erythrocyte-methotrexate	
2	E6TGN	erythrocyte-TGN	
3	ALAT	alanine aminotransferase	https://www.healthline.com/health/alt
4	ASAT	aspartate aminotransferase	https://labtestsonline.org/tests/aspartate-aminotransferase-ast
5	leuk	Leukocyte count	https://en.wikipedia.org/wiki/White_blood_cell
6	lymf	Lymphocyte count	https://en.wikipedia.org/wiki/Lymphocyte
7	neut	Neutrophil count	https://en.wikipedia.org/wiki/Neutrophil
8	trom	Platelet count	https://en.wikipedia.org/wiki/Platelet
9	TGN	thioguanine nucleotides	DOI: 10.1007/s00280-018-3704-7
10	DNA6TGN	thioguanine nucleotides into DNA	DOI: 10.1007/s00280-018-3704-7
11	MMP	Methylated 6-mercaptopurine	https://en.wikipedia.org/wiki/Mercaptopurine
12	TPMT	thiopurine S-methyltransferase	https://en.wikipedia.org/wiki/Thiopurine_methyltransferase
13	MTX	Methotrexate	https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/cancer-drugs/drugs/methotrexate-maxtrex
14	MTXpg1-6	Methotrexate polyglutamates	DOI: 10.1007/s00280-018-3704-7

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.* To provide a deeper understanding of pharmacogenetics/-kinetics/-dynamics of MT and increase cure rates for childhood leukemia. The datasets emerge from two Nordic childhood leukemia protocols (ALL92: 1992-2006) and ALL2008 (2008-2018). [By: Kjeld Schmiegelow]

M2: Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?. Kjeld Schmiegelow and his research lab "Bonkolab" at Rigshospitalet, Copenhagen, Denmark, performed all the MTX/6MP metabolite analyses and the single nucleotide profiling. Treatment centers throughout the Nordic and Baltic region provided clinical data, drug doses, and blood counts. [By: Kjeld Schmiegelow]

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.* The Danish Childhood Cancer Foundation; The Swedish Childhood Cancer Foundation; The Danish Cancer Society; The Nordic Cancer Union; The Novo Nordisk Foundation. [By: Surname, name]

M4: Any other comments? None. [By: Kjeld Schmiegelow]

> 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*

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: Rikke Linnemann Nielsen

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 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

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: Rikke Linnemann Nielsen

MEDICAL AI DATA SHEET

A principled standard for clinical data communication

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]

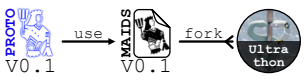
T5: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they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

MAIDS Provenance



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