INF

FACULTY OF
COMPUTER SCIENCE

Triplet-based-learning with the help of crowdlabeling on medical data

Assessing the difficulty of annotating medical data

Anne Rother
Knowledge Management & Discovery Lab

Anne Rother

anne.rother@ovgu.de

10/2017 – present: Business Informatics

04/2018 – present: Student research assistant @ KMD

Papers

PLOS ONE

- Anne Rother, Uli Niemann, Tommy Hielscher, Henry Völzke, Till Ittermann, and Myra Spiliopoulou. Assessing the difficulty of annotating medical data in crowdworking with help of experiments. PLOS ONE, (16)7:1-26, Public Library of Science, July 2021.

frontiers

- Anne Rother, and Myra Spiliopoulou. Virtual Reality for Medical Annotation Tasks- A Systematic Review. Frontiers in Virtual Reality, Frontiers, 2022.

“Rudolf-Kruse-Award“

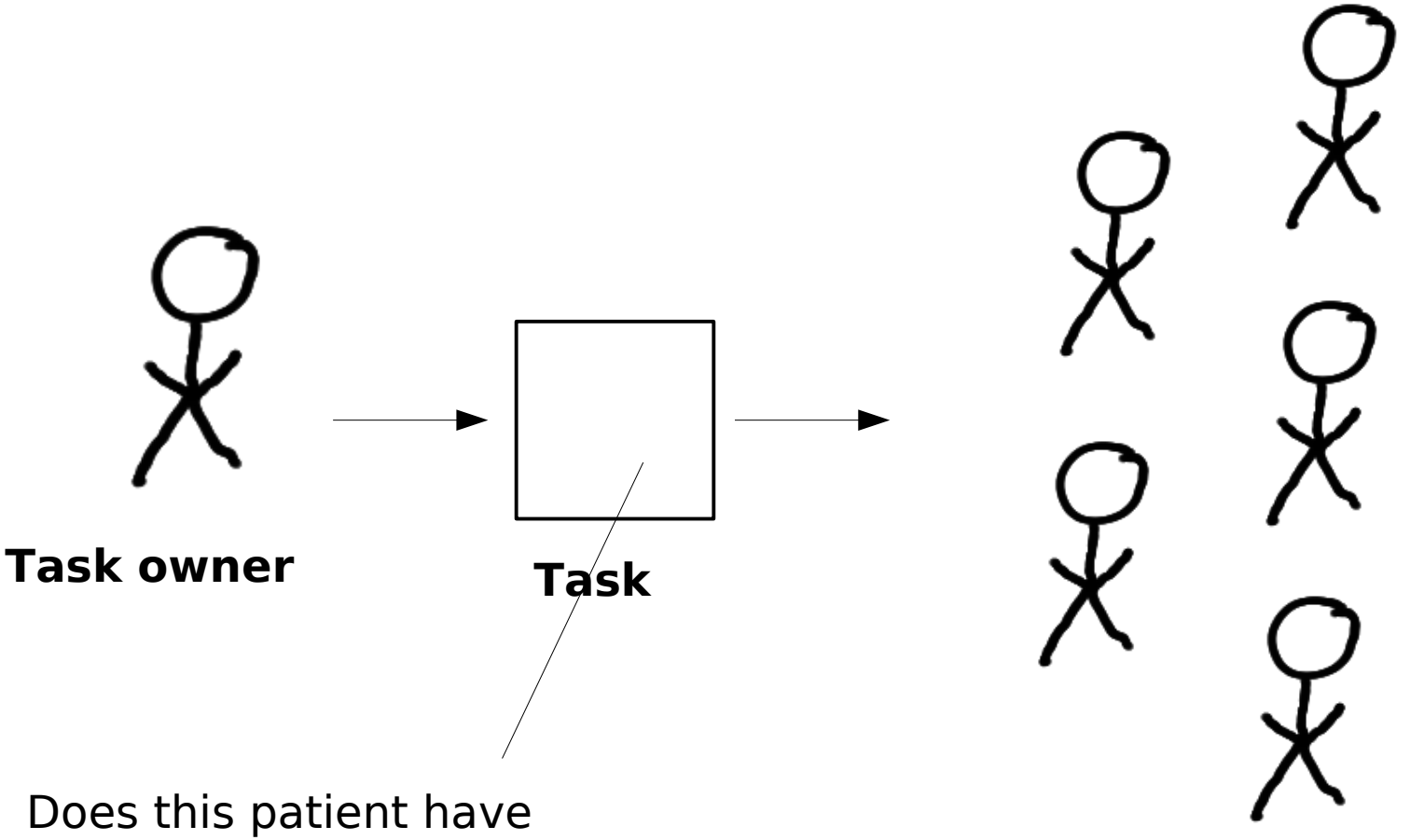
- student research award of the faculty of computer science

Agenda

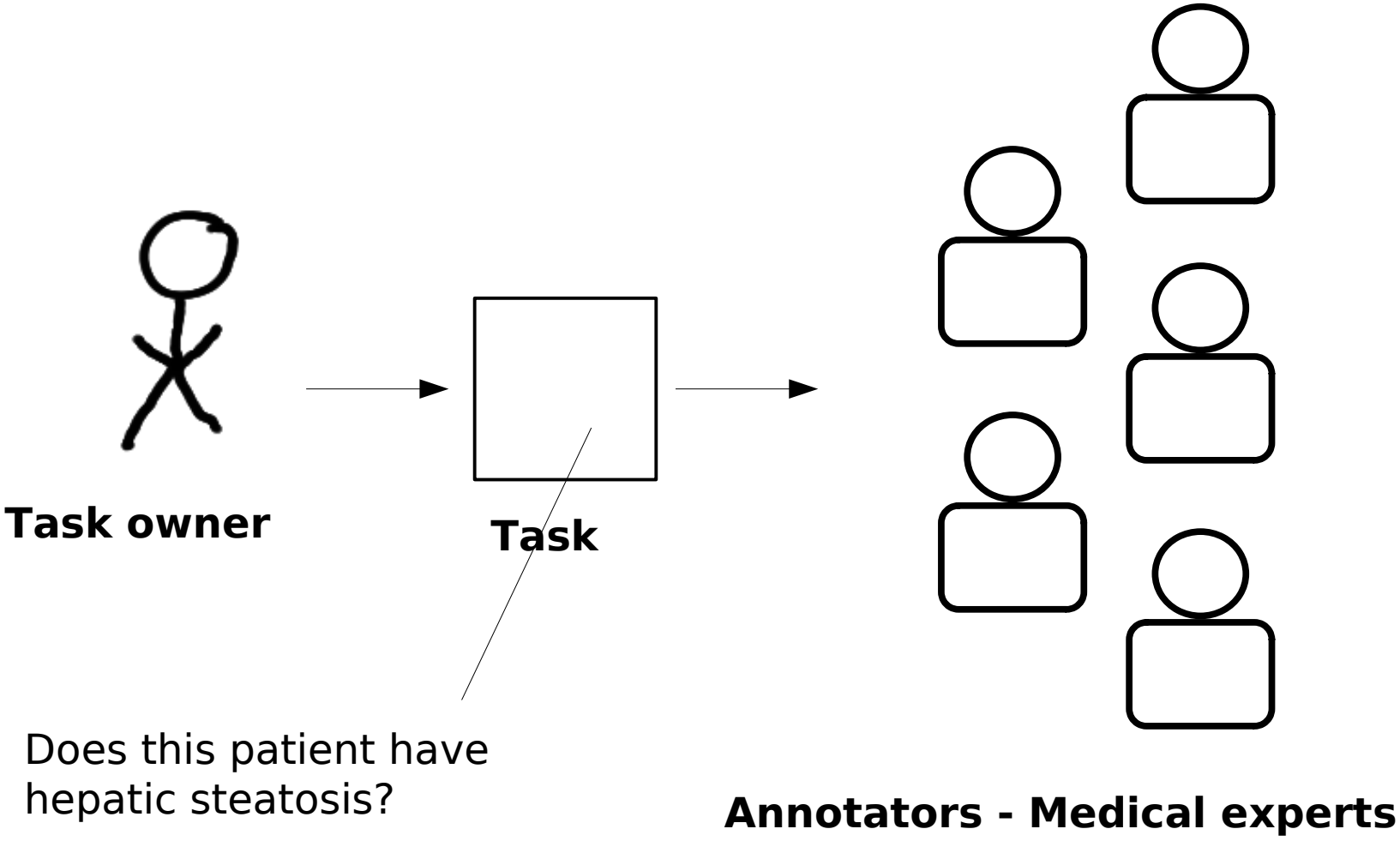
1. Medical crowdlabeling
2. Triplet-based learning
3. Assessing the difficulty of medical data: Example of an experiment

Medical crowdlabeling

1. Medical crowdlabeling



1. Medical crowdlabeling



1. Medical crowdlabeling
“Ask 3 doctors, get 4 opinions” [1]

- Expensive medical-expert labor
- ML solutions that affects patients must have exceptionally high quality
- Minimize uncertainty and misclassification

[1] Quang Duong, Invited Talk “Challenges and Lessons in Adjudicating Labeling Medical Data“ 1ST Data Excellence Workshop, HComp2020.

Triplet-based learning

2. Triplet-based learning

“Object A is more similar to object B than to object C”
d(A, B)<d(A, C) [2]

“Instance A has hepatic steatosis.
Instance C has no hepatic steatosis.
To which instance is B more similar?”

Image taken from [3]

[2] M. Kleindessner and U. von Luxburg, “Kernel functions based on triplet comparisons“ NeurIPS'17, 2017.

[3] M. Wilber, I. Kwak, and S. Belongie, “Cost-Effective HITs for Relative Similarity Comparisons” HCOMP, 2014.

2. Triplet-based learning

There are different application areas e.g.:

- clustering and computing a centroid of a data set [4]
- use kernel function, e.g. compute similarity scores [2]
- improve crowdsourcing tasks – minimize costs [3]

[3] M. Wilber, I. Kwak, and S. Belongie, “Cost-Effective HITs for Relative Similarity Comparisons” HCOMP, 2014.

[2] M. Kleindessner and U. von Luxburg, “Kernel functions based on triplet comparisons“ NeurIPS'17, 2017.

[4] H. Heikinheimo and A. Ukkonen, “The crowd-median algorithm“ HCOMP, 2013.

Assessing the difficulty of medical data

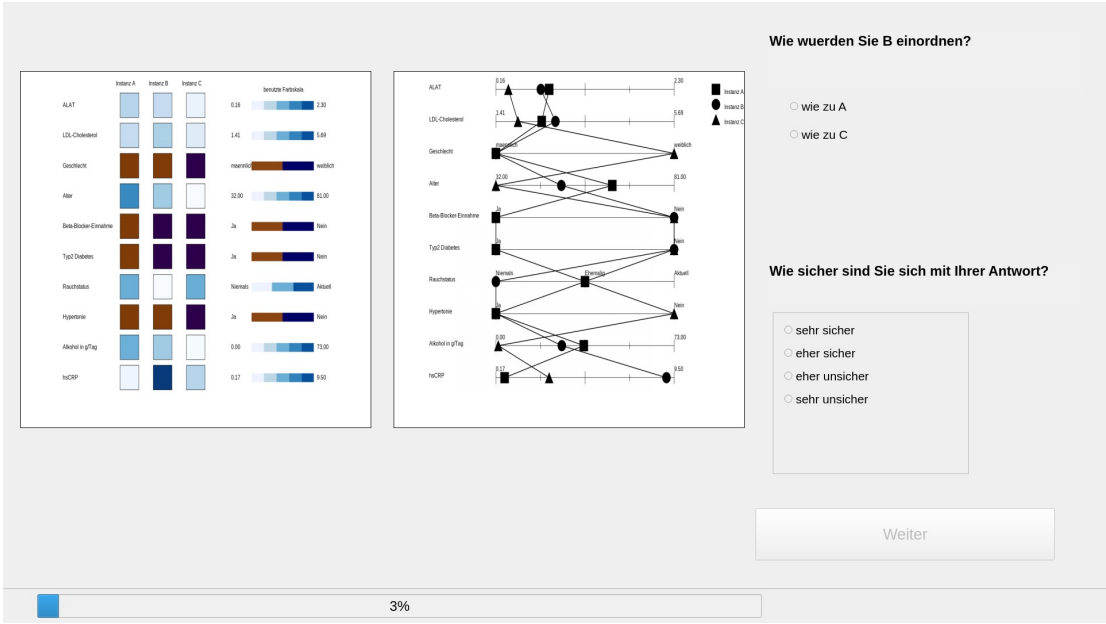
3. Example of an annotation experiment [5]

Task: Identify similarity of participants in a SHIP cohort with respect to a disease

Goal: (1) Detection of the difficulty and influence of the configurations.
(2) Comparison of human and machine

Technology: Visualizations, EDA and activity sensor

Annotators: 29+3 experts



[5] A. Rother, U. Niemann, T. Hielscher, H. Völzke, T. Ittermann, and M. Spiliopoulou, “Assessing the difficulty of annotating medical data in crowdworking with help of experiments”, PLOS ONE, (16)7:1-26, Public Library of Science, July 2021.

3. Assessing the difficulty of medical data

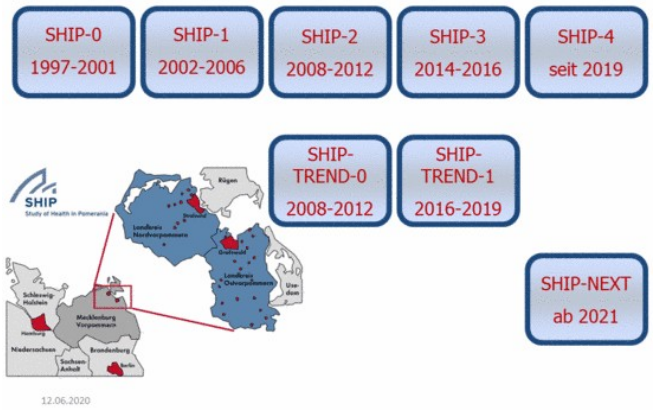
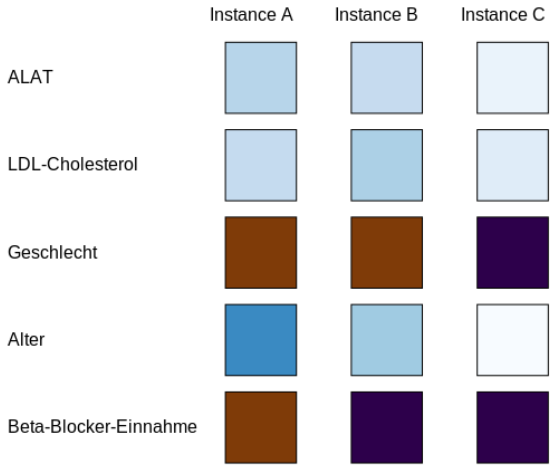


Image taken from [6]

alat_s	ldlch	sex	age	Beta Blocker	Diabetes
0.7	2.84	1	50	No	No
0.31	1.94	2	32	No	No
0.8	2.51	1	64	Yes	Yes
0.2	3.21	2	45	No	No
0.49	5.4	1	52	No	No



[6] "Study of Health in Pomerania" (SHIP) https://www2.medizin.uni-greifswald.de/cm/fv/ship/studienbeschreibung/

3. Assessing the difficulty of medical data

variable	description	values margin
age	age at examination	years
alat_s	alanin-aminotransferase	µkatal/l
alcohol in g/day	alcohol in g/day	g/day
beta blocker	beta blocker intake	yes or no
crp_hs	high-sensitive CRP	mg/l
diabetes	type 2 diabetes mellitus	yes or no
hypertonia	hypertension	yes or no
ldlch	LDL-cholesterol	mmol/l
sex	sex	male or female
smoke_status	smoking status	former, never or current
livfat_per	liver fat concentration	%
stea	hepatis steatosis	0 or 1

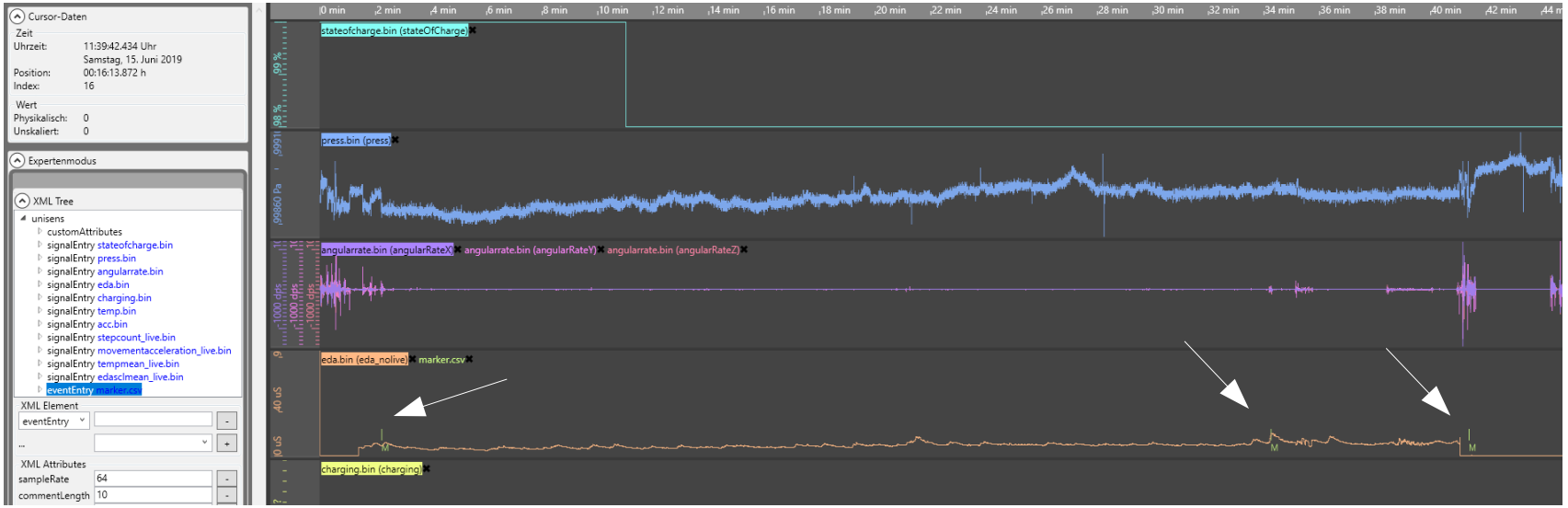
https://doi.org/10.1371/journal.pone.0254764.t001

Overview of used variables in the annotation experiment: The first 10 variables listed in the Table were shown to the experiment participants for each entry. The last two variables were not shown to the experiment participants.

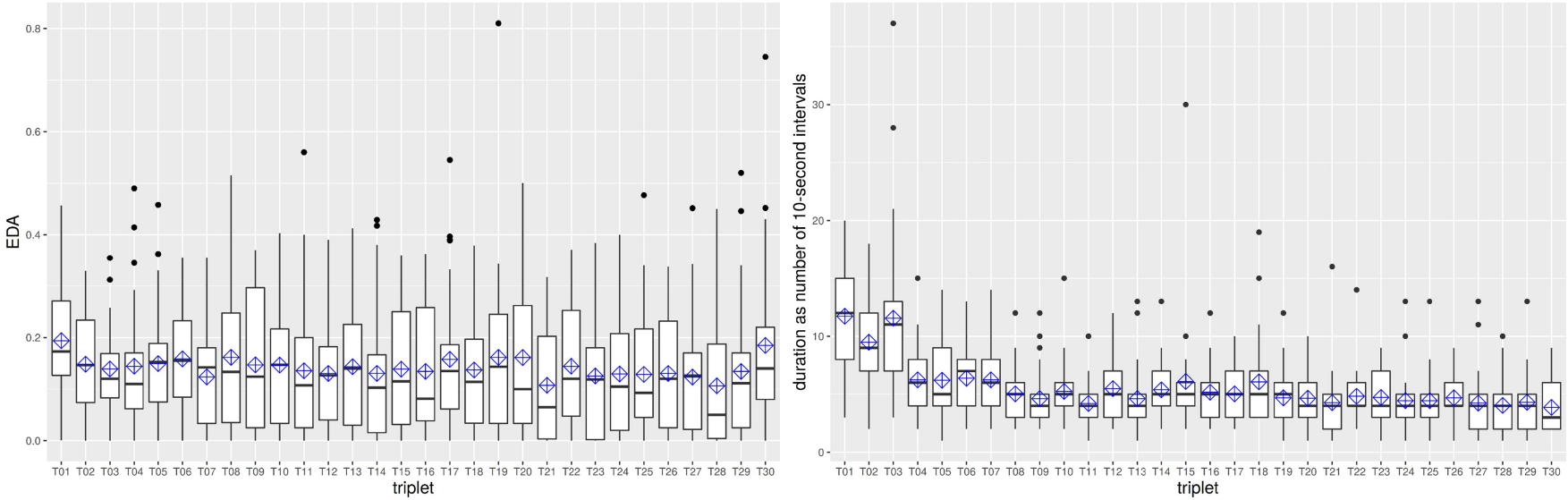
The variable livfat_per depicts the fraction of fat in the liver and was used for the class separation.

Measuring Electrodermal Activity (EDA)

- Measurement and calculation of different paramenters e.g. body position and mean skin conductance in output interval



Task difficulty: EDA and duration

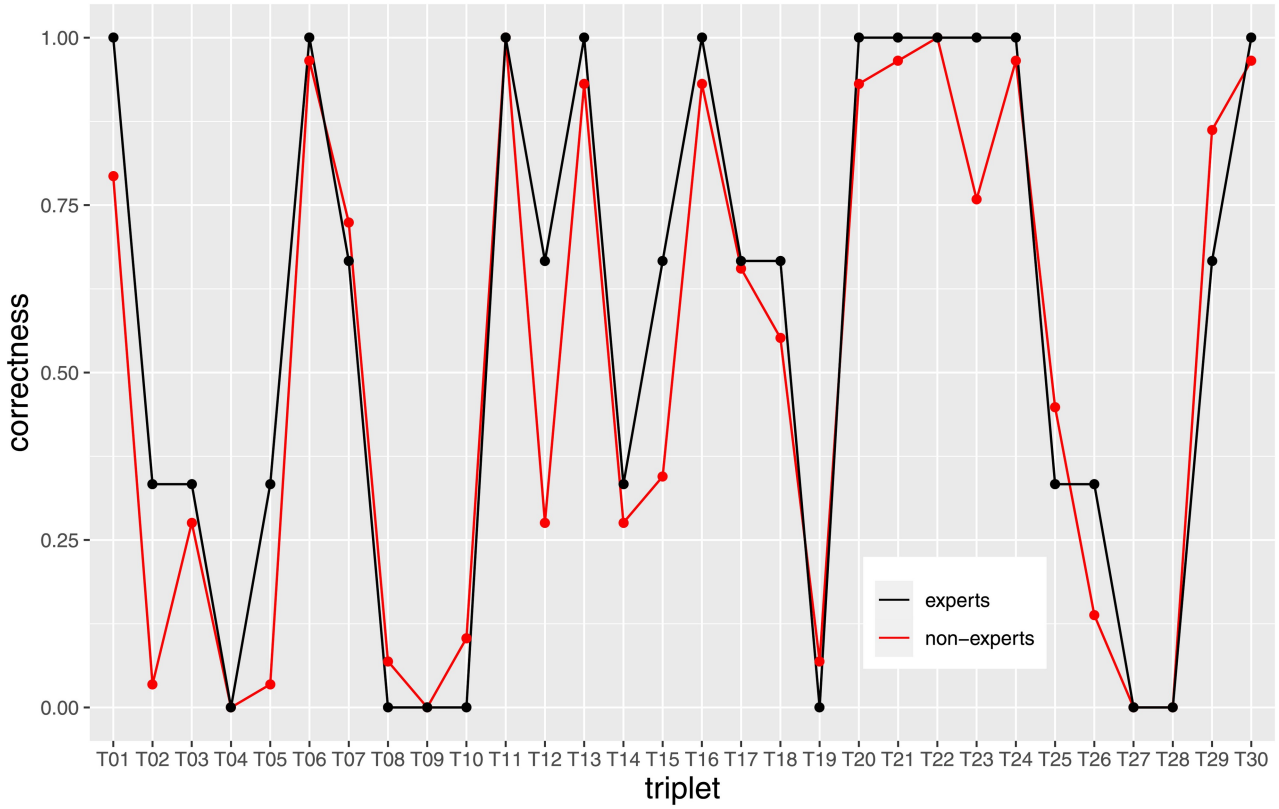


Boxplots of EDA (left subfigure) and of duration (right subfigure) with one box per triplet. The order of the triplets did not have any evident effect on EDA but had an effect on duration.

Task difficulty: correctness - triplets

A slight improvement can be perceived for most of the annotators, the difference is very small, implying that the acclimatization phase has almost no effect on correctness.

Task difficulty: correctness – experts



The two groups performed similarly, and misclassified some triplets in agreement; slight differences can be explained by the large difference between the number of non-experts (29) and the number of experts (3).

Task difficulty: uncertainty, duration, EDA

exposure	outcome	ß (95%-Confidence Interval)	p
Stated_U	duration	6.46 (3.34; 9.58)	<0.001
EDA	duration	9.46 (-7.46; 26.40)	0.273
Stated_U	EDA	-0.00 (-0.01; 0.00)	0.294

https://doi.org/10.1371/journal.pone.0254764.t003

Statistical analysis on the association between stated uncertainty (Stated_U) and duration for each triplet annotation task (first row), between EDA and duration (second row) and between EDA and Stated_U (last row).

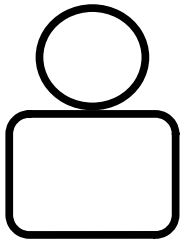
The association between Stated_U and duration is significant.

Results

- No correlation between correctness and either of stated uncertainty, stress and task duration
- Annotator agreement has not been predictive either
- When controlling for Triplet ID, we identified significant correlations, indicating that correctness, stress levels and annotation duration depend on the task itself
- Average correctness among the experiment participants was slightly lower than achieved by ASBA
- Triplet annotation turned to be similarly difficult for experts as for non-experts

Task difficulty: uncertainty, duration, EDA, correctness

non experts



experts

Minimize uncertainty and misclassification

Semi-Supervised Learning on triplets

ASBA

Deterministic mechanism

Closing and outlook

- ✓ Medical crowdlabeling – non experts vs experts
- ✓ Triplet-based-learning
- ✓ Example of an annotation experiment

→ Semi-supervised learning on medical triplets [ongoing]

→ Effect of the number of variables in triplet annotation tasks [ongoing]

Thank you!

Q&A

www.ovgu.de

https://www.kmd.ovgu.de/Research/Learning+in+Dynamic+environments/Crowdsourcing.html