Comparing visual tools for pairwise comparisons of tabular data

Anne Rother[a,*], Matteo Polsinelli[b,0], Till Ittermann[c], Giuseppe Placidi[d] and Myra Spiliopoulou[a]

[a]Otto-von-Guericke University Magdeburg, Magdeburg, Germany
[b]University of Salerno, Salerno, Italy
[c]University Medicine Greifswald, Greifswald, Germany
[d]University of L'Aquila, L'Aquila, Italy
ORCID (Anne Rother): https://orcid.org/0000-0002-6768-5871, ORCID (Matteo Polsinelli): https://orcid.org/0000-0002-4215-2630, ORCID (Till Ittermann): https://orcid.org/0000-0002-0154-7353, ORCID (Giuseppe Placidi): https://orcid.org/0000-0002-4790-4029, ORCID (Myra Spiliopoulou): https://orcid.org/0000-0002-1828-5759

Abstract. AI-based diagnostics demand reliable medical record labeling. Despite the advances of few-shot and zero-shot learning, each specialized medical data collection demands at least some labels that agree with the feature space and the class distribution of the collection. However, human posteriori classification of existing records on diagnoses that have not been considered during the original data acquisition demands effort and expert knowledge. To facilitate human labor and decrease the required level of expertise, we propose a workflow that encompasses pairwise comparisons of medical records and dedicated visualizations for the juxtaposition of record pairs in the original feature space. We evaluate the potential of new visualization schemes in controlled experiments with human volunteers and we juxtapose the results to those achieved with earlier, much simpler visualizations.

1 Introduction

Pairwise comparisons are used in machine learning to derive similarity functions that take *local* proximity between objects into account [20]. Pairwise comparisons are also used crowdworking to capitalize on the fact that humans can discern similarities between objects with their eyes, in a way that AI still cannot immitate [15], [16], [5]. For example, when called to perform a pairwise comparison among the three faces in the upper part of Figure 1, humans are likely to ignore the whiskers, a feature of some importance when comparing the three faces in the lower part of the same figure. When it comes to high-dimensional medical records though, human annotators need more assistance when deciding which features to concentrate on.

In this paper, we investigate the potential of different structured record visualizations in assisting humans in pairwise comparisons. We propose a workflow that encompasses a mechanism for triplet construction from a set of labeled medical records for a binary classification problem (person has the disease: Y/N), two visualizations for pairwise comparisons, an experiment design for the evaluation of these visualizations on volunteers, and a set of evaluation criteria to assess the potential of each method and its merit in comparison to simpler visualization mechanisms.

* Corresponding Author. Email: anne.rother@ovgu.de

Our first contribution is the complete workflow, intended to assist human annotators who do pairwise comparison of structured medical records for the purpose of labeling. Our second contribution consists of the two presented visualizations, which are intended to highlight similarities and differences among records in the original feature space. Our last contribution is the evaluation approach, covering an experiment that involves human volunteers and a retrospective comparison to the results of an earlier experiment that used simpler visualizations.

The paper is organized as follows. We first discuss related work on pairwise comparisons and on visualization of structured medical records, focusing on visualization methods for the original feature space. In section III, we present the elements of our approach, while in section IV we present the medical data we used, the experiment we performed with human volunteers and our evaluation criteria. Section V contains our results and a discussion on them. The last section summarizes the findings and provides an outlook.

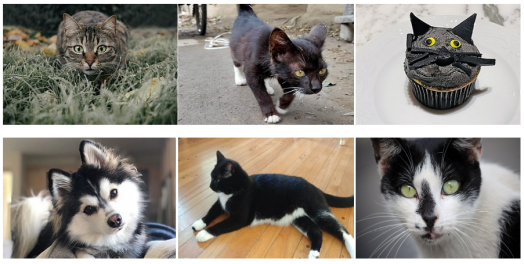

Figure 1. Two pairwise comparisons involving two cats and one muffin (upper part of the figure), respectively one dog (lower part of the figure). Image taken from https://commons.wikimedia.org/wiki/File:Black_and_white_cat_head.jpg

2 Related Work

2.1 Pairwise comparisons

Studied intensively from the machine learning perspective, see e.g. [10], where the objective is to induce a distance function over the

data space. The human-driven process of finding the two most similar objects inside a triplet is investigated in psychology, but there the objective is to acquire insights into human perception [3]. Insights into whether triplet comparisons performed by human annotators are indeed exploitable by machine learning algorithms are mostly limited to the comparison of images [1, 18]. Arguably, pairwise comparison in triplets of tabular data records, such as medical instances, is different from the comparison of image instances. Yao et al. used pairwise comparisons for the estimation of treatment effects in observational data [27]: they chose three pairs of instances, one consisting of the most proximal target instance x_i and control instance x_j, one consisting of the most remote target instance with respect to x_i, and one consisting of the most remote control with respect to x_j. They then introduced two counteracting metrics on the basis of loss functions, intended to bring similar instances close to each other but not too close in the representation space.

2.2 *Measuring the difficulty of annotation and labeling tasks*

Difficulty of pairwise comparisons of images has been investigated in [1, 2]. Similarly to our earlier works [7, 15] on pairwise comparisons of non-image data. Ahonen et al. [1] used sensors that measure electrodermal activity. Their results were not conclusive, in the sense that it did not become evident what makes a comparison difficult independently of the person who performs the comparison. The difficulty of pairwise comparisons of non-image objects is less investigated in general, despite the fact that non-image objects are of relevance in several application domains, including the annotation of clinical data. However, there are several investigations on the difficulty of crowdworkers tasks, including labeling tasks and more elaborate annotations. Traditionally, 'difficulty' (which is not observable) is modeled on the basis of observable quantities. One of them is 'duration', defined in [4] as the time needed to complete a specific task and used as indicator of task difficulty for a specific crowdworker. An important indicator is (dis)agreement among crowdworkers, pointing to task ambiguity [17] or to diverging interpretations of a task [9], i.e. to inherent task properties independently of a specific crowdworker's skills and expertise. In [15] we focused on (dis)agreement as potential indicator of difficulty: Annotator (dis)agreement was not predictive – neither for difficulty nor for correctness. Furthermore annotators performed pairwise comparisons on triplets that consisted of 10-dimensional medical instances from the cohort SHIP-2 of [24]. We found that for some instances proximity across certain dimensions was misleading in the sense that annotators consistently decided that a pair of instances inside a triplet were more similar than they truly were.

2.3 *Annotation of medical data*

Images, diagnostic texts or structured instances, is a very important task, for which crowd-working has been applied increasingly and successfully in recent years [22, 25, 26]. In [26], Wazny et al. list 8 areas of medical applications, where crowdsourcing is being used; among them, diagnosis, such as assigning scores to tumors. This corresponds to the creation of ground truth in existing datasets through labeling. However, medical annotations go beyond the assignment of labels or scores. For example, Joshi et al. recruited volunteers who identified the 'location' of emotional episodes in timestamped data, as well as the duration of these episodes [8]. Studies on the annotation of medical data follow different directions. They include the study of the potential of Virtual Reality (VR) technologies as in [6, 13], the generation of open access datasets [11], the role of annotated data collections in education [23], and ways of semi-automating the labeling/annotation process. Among the latter, the earlier work of Nissim et al. [14] highlighted the potential of active learning to reduce label acquisition cost. More recently, combinations of semi-supervision and crowdsourcing have become a popular subject of investigation, see e.g. [19, 21].

3 Workflow for record annotation through pairwise visual comparisons

3.1 *A pie-based visualization*

The proposed method was inspired by the solution proposed in [15] in which the experiment participant was shown two representations: a tile-based and a line-based. In the first, each triplet is composed of ten tiles for each risk factor with the numerical values marked as shade (Figure 2, left box). In the second, the position of the middle record value for some variables indicates its distance from the variable values for the other two records (Figure 2, right box). This solution has been shown to be effective but can be improved using a new visualization method that does not separate the features from the others.

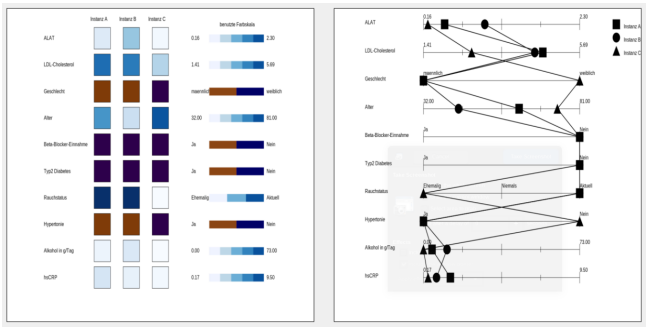

Figure 2. Triplet for low level of difficulty: Old visualization based on the experiment from [15]

The main idea is to use the pie-based visualization shown in Figure 3. Compared to the old visualization, this is more compact since each of the ten variables is represented as a slice of the pie. In this way, three pies are necessary to represent the three subjects A, B, and C of the experiment. The comparison between subjects is immediate, and the slices of the pie are position invariant, since the crowd worker is not biased by the particular arrangement of each variable (there is no ordering between them).

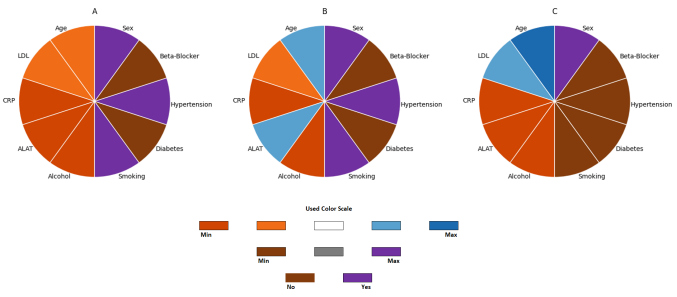

Figure 3. Triplet for low level of difficulty: New visualization

In both methods, the color palette is assigned by linearly distributing the colors in the Min-Max interval of the feature values by using a discrete number of colors for discrete features. 5-values color scales is used for continuous features, the 2-values color scales is used for binary features, and the 3-values color scales is used for the ternary feature.

The resulting color-based triplet assignments are described in Algorithm 1 for the old method and in Algorithm 2 for the new method.

Algorithm number 1 tripletA, tripletB, tripletC

$feat \in Features\ m \leftarrow min(feat)\ M \leftarrow max(feat)$

$feat\ is\ binary\ bins \leftarrow 2\ feat\ is\ ternary\ bins \leftarrow 3\ bins \leftarrow 5$

Palette $\leftarrow$ createPalette(bins, m, M) $tripletA(feat) \leftarrow closer(Palette, tripletA.feat.val)$ $tripletB(feat) \leftarrow closer(Palette, tripletB.feat.val)$ $tripletC(feat) \leftarrow closer(Palette, tripletC.feat.val)\ show(Palette)$

Algorithm number 1 tripletA, tripletB, tripletC

$nFeature \leftarrow 10\ pieA \leftarrow createpie(nFeature)\ pieB \leftarrow createpie(nFeature)\ pieC \leftarrow createpie(nFeature)\ k \leftarrow 1$

$F \in Features\ m \leftarrow min(F)\ M \leftarrow max(F)$

$F\ is\ binary\ bins \leftarrow 2\ F\ is\ ternary\ bins \leftarrow 3\ bins \leftarrow 5$

$Palette \leftarrow createPalette(bins, m, M)\ pieA(k) \leftarrow closer(Palette, tripletA.feat.val)\ pieB(k) \leftarrow closer(Palette, tripletB.feat.val)\ pieC(k) \leftarrow closer(Palette, tripletC.feat.val)\ k \leftarrow k + 1\ Palette \leftarrow createPalette(5, 0, 1)\ show(Palette)$

In Figure 3, it is possible to understand how easily it can be concluded that instance B is similar to instance A because the right half-pie of both is equal, as well as the slices representing LDL, CRP and Alcohol. Instead, in Figure 2, in which the same instances A,B,C are represented, the comparison is less immediate because the crowd-worker is led to analyze one variable at a time.

This is even more evident in Figures 4 and 5 which present a less obvious case. In fact, instance B is still more similar to instance A, but in this case, the similarities are few and it is not possible to establish it by directly confronting each variable, but it is necessary an overall view, and for this reason, pie-based visualization is still superior.

The last example is presented in Figure 6 and Figure 7 and is very difficult to assess. Both instances A and C are good candidates and looking carefully in the pie-based visualization, it is possible to conclude that B is more similar to A, even if even if very little.

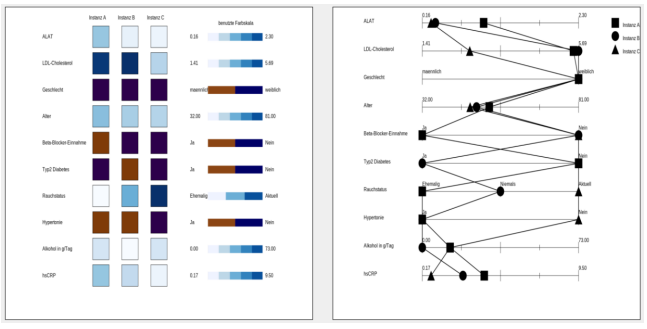

Figure 4. Triplet for middle level of difficulty: Old visualization based on the experiment from [15]

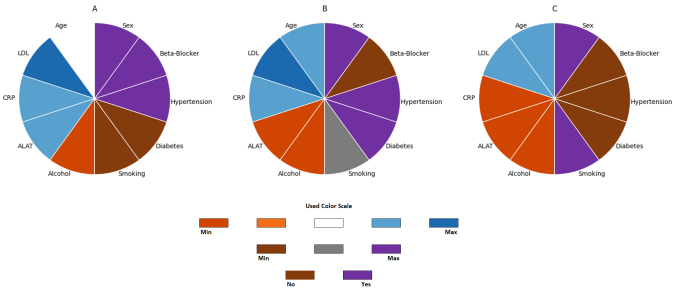

Figure 5. Triplet for middle level of difficulty: New visualization

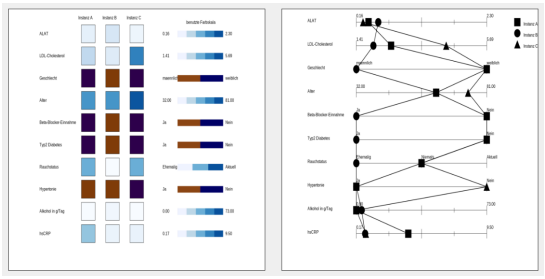

Figure 6. Triplet for high level of difficulty: Old visualization based on the experiment from [15]

4 Our Evaluation Workflow

4.1 The triplets of the experiment

In this study, we investigate the potential of different visualization schemes for pairwise comparison of medical records.

As a follow-up of the experiment in [15] we asked 2 experts to annotate the new visualization to assess whether an individual is more similar to a healthy or a diseased individual using hepatic steatosis as an outcome. Both experts conduct research on active learning, prediction and classification. They do not know the SHIP dataset. Each expert was asked to annotate 30 annotation tasks + 3 tasks of different levels of difficulty. Furthermore, they have to express the perceived difficulty for the annotation of each triplet by choosing one of the following four answers: “very certain", “rather certain", “rather uncertain", and “very uncertain".

For choosing the triplets we used the dataset as presented and described in [15]. There we randomly selected 90 records out of 852 individuals of SHIP-2. These are categorized into the following three categories: “no hepatic steatosis" (liver fat fraction $\leq$ 5.0%, n = 501), “mild hepatic steatosis" (5.0% $\leq$ liver fat fraction <14%, n = 238), and “moderate to severe hepatic steatosis" (liver fat fraction $\geq$ 14%,

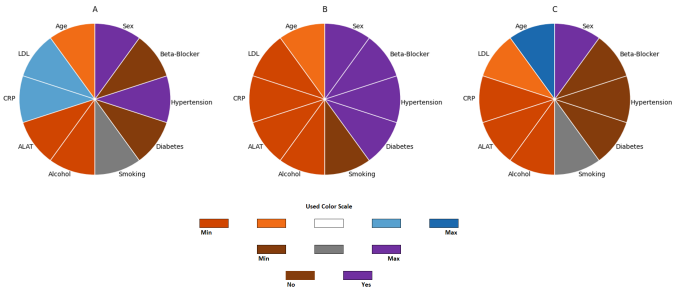

Figure 7. Triplet for high level of difficulty: New visualization

n = 113) [12]. More specific we selected 45 individuals from the class “no hepatic steatosis" and 45 from the class “moderate to severe hepatic steatosis" and split this two subsamples into three groups of 15 individuals. For each subject, ten risk factors of hepatic steatosis are reported: age, sex, alanine-aminotransferase (ALAT), low-density lipoproteine (LDL) cholesterol, alcohol consumption, hypertension, beta-blocker intake, type 2 diabetes mellitus, smoking status, and c-reactive protein (CRP).

4.2 Evaluation Criteria

Our scenario is a controlled pairwise comparison experiment, in which we want to find out which features catch the participants' eye under each configuration and which configuration helps them most in finding the ‘good features’. The configurations are (a) our new color-based one and (b) the baseline used in the article of [15].

To compare the new graphic model with the article of [15], we compute correctness, and then we run the experiment with the same triplets. We compute the average correctness as performance indicators to evaluate the new graphic model for different degrees of task difficulty.

To evaluate the performance of both methods, we define the following evaluation criteria:

- Correct classifications
- Score, to compare the two visualizations: How often one was correct under each visualization

In addition, we present the uncertainty of experts for the new visualization.

5 Findings

5.1 Findings with the proposed visualization

In Table 1 we show the annotation of the two experts. They differ in the annotation in 6 tasks (bold marked). Furthermore, the column “Uncertainty" shows the perceived difficulty per triplet.

As depicted both experts are “rather certain" in the annotation: 14 and 12 times out of 30. “Rather uncertain" they are in 9 and 10 triplets out of 30. On 4 and 6 triplets, they are “very uncertain". The experts gave the lowest response for “very certain": Only 3 and 2 times out of 30 triplets they chose this answer.

It is remarkable that the annotation of the triplets for easy, medium and difficult differ. T11 represent the easy triplet - here the difficulty changes slightly. For middle difficulty the annotation changes completely. Under T18, both experts annotated incorrectly. Later on, they annotated correctly when they annotated this task again. For the difficulty triplet the perceived difficulty changes from very uncertain to rather uncertain. The annotation remains the same, but incorrect.

5.2 Comparison to the baseline visualization

In Table 2 we juxtaposed how the experts annotated the triplets for both visualizations. For better comparability we removed one expert annotation for the old version. This expert is an epidemiologist and created the dataset.

The annotations differ in 14 out of 30 tasks and are marked in bold. The new visualization was annotated slightly better than the old visualization. We have better correctness for the easy triplets, similar correctness for the medium ones, and also similar for the difficult ones. On average, the old visualization was correctly annotated 0.50,

	Correctness		Uncertainty	
Triplet	Expert 1	Expert 2	Expert 1	Expert 2
T01	**yes**	**no**	rather uncertain	rather certain
T02	yes	yes	rather uncertain	rather uncertain
T03	yes	yes	rather certain	rather uncertain
T04	no	no	very certain	very certain
T05	yes	yes	rather certain	rather certain
T06	yes	yes	rather uncertain	rather uncertain
T07	**yes**	**no**	rather uncertain	rather uncertain
T08	no	no	rather certain	rather certain
T09	no	no	very certain	very certain
T10	yes	yes	very certain	rather uncertain
T11	yes	yes	rather certain	rather certain
T12	no	no	rather uncertain	very uncertain
T13	yes	yes	rather certain	rather certain
T14	no	no	rather certain	rather certain
T15	no	no	rather certain	rather certain
T16	yes	yes	very uncertain	rather uncertain
T17	yes	yes	rather certain	very uncertain
T18	yes	yes	very uncertain	very uncertain
T19	**yes**	**no**	rather certain	rather certain
T20	no	no	rather uncertain	rather certain
T21	yes	yes	rather uncertain	rather certain
T22	yes	yes	rather certain	rather uncertain
T23	**no**	**yes**	rather uncertain	rather uncertain
T24	yes	yes	rather uncertain	rather certain
T25	**no**	**yes**	rather certain	rather certain
T26	yes	yes	very uncertain	very uncertain
T27	no	no	rather certain	rather uncertain
T28	no	no	rather uncertain	very uncertain
T29	**no**	**yes**	rather certain	rather certain
T30	no	no	very uncertain	very uncertain
T31	yes	yes	very uncertain	rather certain
T32	no	no	rather uncertain	very uncertain
T33	no	no	rather certain	rather certain

Table 1. Results of the expert-annotation for each triplet. T31, T32 and T33 represent the triplets chosen for easy, medium and difficult.

the new visualization on average 0.57. This could also be related to the choice of experts. In the old visualization, a physician annotated the triplets and another expert knew the SHIP-2 dataset. In contrast, the two new experts for the new visualization have no medical background and do not know the data set. We are not trying to find the most globally influential variable. Since the important variables vary per triplet. Therefore, each variable has the same position in each triplet.

6 Conclusion and Future Work

In this work, we investigated the potential of different visualization schemes of medical records. We elaborated on an experiment whether a new visualization leads to a better annotation, based on correctness and investigated this with expert annotation on a previous visualization. Thereafter, we will start investigating the role of stress as a confounder. We will also expand the experiment to non-experts and focus on uncertainty, to further improve the visualization and thus get better results in the annotation. Moreover, we will investigate which features are affecting correctness and how to combine with semisupervised pairwise comparisons.

6.1 Further possibilities for data annotation

In addition to various visualization methods, annotation can also take place on the basis of raw data, for example as tabular data (see Table 3). Table 3 shows a simple triplet. The middle, B, instance is

Triplet	Correctness under the old visualization	Correctness under the new visualization
T01	1	0.5
T02	0.5	1
T03	0	1
T04	0	0
T05	0.5	1
T06	1	1
T07	0.5	0.5
T08	0	0
T09	0	0
T10	0	1
T11	1	1
T12	0.5	0
T13	1	1
T14	0	0
T15	0.5	0
T16	1	1
T17	0.5	1
T18	0.5	1
T19	0	0.5
T20	1	0
T21	1	1
T22	1	1
T23	1	0.5
T24	1	1
T25	0.5	0.5
T26	0.5	1
T27	0	0
T28	0	0
T29	0.5	0.5
T30	1	0
Average	0.50	0.57

Table 2. Overview of the annotation of each triplet for both Visualizations: 1 means 'both experts assigned the middle instance correctly', 0 means 'both experts assigned the middle instance incorrectly' and 0.5 means 'the two experts disagreed'.

to be assigned whether it is more similar to the A instance or C instance. Similar variables are marked in blue (B more similar to A) or orange (B more similar to C). In this example, the IRIS data set consists of only a few variables, so that a more manageable assessment can be made. In this example, annotators would look at how many matches there are per variable (the class is not visible) and then decide whether the B instance is more similar to the A instance or to the C instance. A rather more difficult example is in Table 4. This is also based on the IRIS data set, but the assignment is made more difficult by the similarity of the A and C instances. The variable "sepal lengh" is not unique in this example. Annotators could therefore possibly ignore this variable for the decision-making process. In the triplet as a whole, the B instance is slightly more similar to the C instance than to the A instance. As soon as a variable is weighted more importantly, this decision could either strengthen the decision or lead to a different decision. With data sets that contain more variables, such as the mushroom data set, it is very difficult to recognize individual variables separately. Our suggestion would be to hide the variables where the values are identical so that a better assignment can take place. This and the optimal number of variables per triplet will be investigated in future experiments.

Funding

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, supported by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania.

instance	sepal length	sepal width	petal length	petal width	class
A	4.7	3.2	1.3	0.2	Iris-setosa
B	**4.6**	**3.1**	**1.5**	**0.2**	Iris-setosa
C	6.1	3.0	4.6	1.4	Iris-versicolor

Table 3. easy triplet based on iris dataset

instance	sepal length	sepal width	petal length	petal width	class
A	5.0	3.4	1.5	0.2	Iris-setosa
B	**6.4**	**2.9**	**4.3**	**1.3**	Iris-versicolor
C	5.1	2.5	3.0	1.1	Iris-versicolor

Table 4. difficult triplet based on iris dataset

References

[1] L. Ahonen, B. Cowley, J. Torniainen, A. Ukkonen, A. Vihavainen, and K. Puolamaki. S1: Analysis of electrodermal activity recordings in pair programming from 2 dyads. *PLoS One. Retrieved from http://journals. plos. org/plosone/article/asset*, 2016.

[2] E. Amid and A. Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, pages 1472–1480, 2015.

[3] N. Diersch, J. P. Valdes-Herrera, C. Tempelmann, and T. Wolbers. Increased hippocampal excitability and altered learning dynamics mediate cognitive mapping deficits in human aging. *Journal of Neuroscience*, 41 (14):3204–3221, 2021.

[4] U. Gadiraju, G. Demartini, R. Kawase, and S. Dietze. Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW)*, 28 (5):815–841, 2019.

[5] A. Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

[6] A. Huaulmé, F. Despinoy, S. A. H. Perez, K. Harada, M. Mitsuishi, and P. Jannin. Automatic annotation of surgical activities using virtual reality environments. *International journal of computer assisted radiology and surgery*, 14(10):1663–1671, 2019.

[7] N. Jambigi, T. Chanda, V. Unnikrishnan, and M. Spiliopoulou. Assessing the difficulty of labelling an instance in crowdworking. In *2nd Workshop on Evaluation and Experimental Design in Data Mining and Machine Learning@ ECML PKDD 2020*, 2020.

[8] A. A. Joshi, M. Chong, J. Li, S. Choi, and R. M. Leahy. Are you thinking what i'm thinking? synchronization of resting fmri time-series across subjects. *NeuroImage*, 172:740–752, 2018.

[9] S. Kairam and J. Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648, 2016.

[10] M. Kleindessner and U. von Luxburg. Kernel functions based on triplet comparisons. In *Advances in neural information processing systems*, pages 6807–6817, 2017.

[11] E. E. Kpokiri, R. John, D. Wu, N. Fongwen, J. Z. Budak, C. C. Chang, J. J. Ong, and J. D. Tucker. Crowdsourcing to develop open-access learning resources on antimicrobial resistance. *BMC infectious diseases*, 21(1):1–7, 2021.

[12] J.-P. Kühn, D. Hernando, A. Muñoz del Rio, M. Evert, S. Kannengiesser, H. Völzke, B. Mensel, R. Puls, N. Hosten, and S. B. Reeder. Effect of multipeak spectral modeling of fat for liver iron and fat quantification: correlation of biopsy with mr imaging results. *Radiology*, 265 (1):133–142, 2012.

[13] O. Legetth, J. Rodhe, S. Lang, P. Dhapola, M. Wallergård, and S. Soneji. Cellexalvr: A virtual reality platform to visualize and analyze single-cell omics data. *Iscience*, page 103251, 2021.

[14] N. Nissim, M. R. Boland, N. P. Tatonetti, Y. Elovici, G. Hripcsak, Y. Shahar, and R. Moskovitch. Improving condition severity classification with an efficient active learning based framework. *Journal of biomedical informatics*, 61:44–54, 2016.

[15] 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):e0254764, 2021.

[16] A. Rother, T. Ittermann, and M. Spiliopoulou. Semi-supervised learning with pairwise instance comparisons for medical instance classification. In *International Symposium on Intelligent Data Analysis*. Springer, 2025. to appear.

[17] M. Schaekermann, E. Law, K. Larson, and A. Lim. Expert disagreement in sequential labeling: A case study on adjudication in medical time series analysis. In *SAD/CrowdBias@ HCOMP*, pages 55–66, 2018.

[18] S. Sharifi Noorian, S. Qiu, U. Gadiraju, J. Yang, and A. Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, pages 882–892, 2022.

[19] W. Shi, V. S. Sheng, X. Li, and B. Gu. Semi-supervised multi-label learning from crowds via deep sequential generative model. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1141–1149, 2020.

[20] V. Simard, M. Rönnqvist, L. Lebel, and N. Lehoux. A method to classify data quality for decision making under uncertainty. *ACM Journal of Data and Information Quality*, 2022.

[21] P. A. Traganitis and G. B. Giannakis. Bayesian semi-supervised crowdsourcing. *arXiv preprint arXiv:2012.11048*, 2020.

[22] J. D. Tucker, S. Day, W. Tang, and B. Bayus. Crowdsourcing in medical research: concepts and applications. *PeerJ*, 7:e6762, 2019.

[23] M. van Deursen, L. Reuvers, J. D. Duits, G. de Jong, M. van den Hurk, and D. Henssen. Virtual reality and annotated radiological data as effective and motivating tools to help social sciences students learn neuroanatomy. *Scientific Reports*, 11(1):1–10, 2021.

[24] H. Völzke, J. Schössow, C. O. Schmidt, C. Jürgens, A. Richter, A. Werner, N. Werner, D. Radke, A. Teumer, T. Ittermann, et al. Cohort profile update: The study of health in pomerania (ship). *International journal of epidemiology*, 2022.

[25] C. Wang, L. Han, G. Stein, S. Day, C. Bien-Gund, A. Mathews, J. J. Ong, P.-Z. Zhao, S.-F. Wei, J. Walker, et al. Crowdsourcing in health and medical research: a systematic review. *Infectious diseases of poverty*, 9 (1):1–9, 2020.

[26] K. Wazny. Applications of crowdsourcing in health: an overview. *Journal of global health*, 8(1), 2018.

[27] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31:2633–2643, 2018.