Crowd-Based Assessment of Deformational Cranial Asymmetries

Kathrin Borchert, Matthias Hirth, Angelika Stellzig-Eisenhauer, Felix Kunz

Crowd-based Assessment of Deformational Cranial Asymmetries

Kathrin Borchert[1], Matthias Hirth[2], Angelika Stellzig-Eisenhauer[3], and Felix Kunz[3]

[1] University of Würzburg, Würzburg, Germany
kathrin.borchert@informatik.uni-wuerzburg.de
[2] TU Ilmenau, Ilmenau, Germany
matthias.hirth@tu-ilmenau.de
[3] University Hospital Würzburg, Würzburg, Germany
[stellzig_a | kunz_f]@ukw.de

Abstract. Crowdsourcing allows collecting subjective user ratings promptly and on a large scale. This enables, for example, building subjective models for the perception of technical systems in the field of quality of experience research or researching cultural aspects of the aesthetic appeal. In addition to research in technical domains, crowdsourced subjective ratings also gain more and more relevance in medical research, like the evaluation of aesthetic surgeries. In line with this, we illustrate a novel use-case for crowdsourced subjective ratings of deformational cranial asymmetries of newborns. Deformational cranial asymmetries are deformations of a newborn's head that might, e.g., result from resting on the same spot for a longer time.
Even if there are objective metrics to assess the deformation objectively, there is only a little understanding of how those values match the severity of the deformational cranial asymmetries as *subjectively perceived* by humans. This paper starts filling this gap by illustrating a crowdsourcing-based solution to collect a large set of subjective ratings on examples of deformational cranial asymmetries from different groups that might have a different perception of those deformations. In particular, we consider pediatricians, parents of children with cranial deformation, and naive crowdworkers. For those groups, we further analyze the consistency of their subjective ratings, the differences of the ratings between the groups, and the effects of the study design.

Keywords: crowdsourcing · subjective assessment · medical data · deformational cranial asymmetries.

1 Introduction

Crowdsourcing gives easy and cost-effective access to a large and diverse group of people. Therefore, crowdsourcing has become an established tool to acquire participants for surveys and user studies. Besides collecting objective information, e.g., shopping behavior, crowdsourcing surveys are also often used for collecting

subjective ratings, e.g., to investigate the quality of technical systems from the user's perspective or to analyze differences in the perception of the aesthetic appeal depending on the cultural background of participants.

One major field of application for those large-scale subjective surveys is quality of experience (QoE) [6] research that targets at understanding, modeling, and optimizing the user's perceived quality of a technical system. In addition to research in technical domains, the usage of crowdsourced subjective studies also gains more and more relevance in medical research, like the evaluation of aesthetic surgeries. Today, objective measurements and metrics for biometric data are well studied and discussed. Therefore, disease patterns can be objectively classified and quantified. However, besides this objective perspective, human perception also needs to be considered in the evaluation of the outcome of treatments. One illustrative example of this are deformational cranial asymmetries. Deformational cranial asymmetries are a deformation of newborn's head caused by always resting on the same spot, for example. This leads to a flattening of the shape of the head as the head is malleable during the first month after birth. While the deformations can be quantified using modern 3D scanners and several objective metrics, no commonly agree thresholds for those metrics exist when to start or stop therapies. Further, it remains unclear when a head is *perceived* to be asymmetric by the general public.

Large-scale online user studies, similar to existing works in QoE research, can help to solve these open questions. Still, running those studies to analyze the perception of experts and non-experts with medical data, leads to new challenges including privacy issues due to the sensitivity of the data or new challenges while displaying the data online, due to the unique data formats used for storing medical data or their pure size.

In this work, we introduce a novel medical use-case for the collection of subjective ratings, namely the large scale subjective assessment of deformational cranial asymmetries. We develop an online user study that displays complex medical data and still preserves the patients' privacy. With the help of this tool, we collect assessments of the deformation severity from different groups of participants. Based on the collected data, we evaluate if the perception varies between people with different background and knowledge about deformational cranial asymmetries. In detail, the ratings of pediatricians, other physicians, laypersons including crowdworkers and non-crowdworkers as well as affected persons, i.e., parents of children with deformational cranial asymmetries, are analyzed and compared. Further, the impact of the study design on the ratings is evaluated.

The remainder of this work is structured as followed. Section 2 provides the background of deformational cranial asymmetries and the objective measurements used to quantify the deformations. Further, an overview of the usage of crowdsourcing in medical research, especially for collecting subjective ratings, is given. Section 3 describes the data set used in our study and details on the study design. The evaluation of the study results is given in Section 4. Section 5 concludes this paper.

2 Background and Related Work

In this section, we first provide background about deformational cranial asymmetries, then give an overview of the usage of crowdsourcing for medical research, especially about the collection of subjective ratings.

2.1 Deformational Cranial Asymmetries

The head of a newborn is malleable, and therefore its shape is deformable, e.g., by resting on the same spot over a long time or due to prenatal reasons. Such deformations are also known as deformational cranial asymmetries [12]. If the deformation is more advanced, a therapy for aesthetic and medical reasons is necessary. There exists several objective metrics to classify [2] and to quantify the severity of the deformation, e.g., including biometrical information like characteristics of the neck muscles [3]. In this work, we use 3D-stereophotogrammetric scans obtained by using the methodology introduced by Meyer-Marcotty et al. [9]. Here, the non-invasive 3D scans are created with a special scanner[4]. By using the software Cranioform Analytics[5], metrics about the shape and volume of the head are determined. These metrics include, e.g., the Cephalic Index (CI) which defines the ratio of the maximum width to the maximum length of a head, the ear shift as well as the anterior and posterior cranial asymmetry index (ACAI/PCAI). The indexes ACAI and PCAI represent the ratio of the volumes of different quadrants of the head.

However, even if there are objective metrics to quantify the deformation, these are not fixed thresholds when to start or to stop the medical treatment [19]. This is mainly because the subjective perception of the grade of deformation is not fully understood yet and may even differ between experts, e.g., physicians, and affected people, e.g., parents of newborns with deformational cranial asymmetries, as well as non-experts. Crowdsourcing is one possibility to acquire a large number of subjective assessments for deformational cranial asymmetries from a diverse set of participants. These assessments can then help to gain an understanding of the perception of the deformations and ultimately be used to derive guidelines for therapies.

2.2 Crowdsourcing and Medical Research

There is a large community focusing on the usage of crowdsourcing in the context of medical research [13,17,10]. The fields of application are ranging from the area of machine learning, e.g., labeling medical big data [15] or improving automatic speech recognizer [14], to recruiting participants for medical user studies [11,5]. Further, the crowdworkers could also provide medical diagnosis [1]. For example, the work of Meyer et al. [8] introduces the platform CrowdMed that gives

[4] `http://www.3dmd.com/` Accessed Jun. 2019

[5] `https://www.cranioform.de/fuer-aerzte/medizinische-informationen.html` Accessed Jun. 2019

crowdworkers access to data of patients with undiagnosed illnesses and Li et al. [7] focusing on the reliability of such crowdsourced diagnosis.

Besides the collection of diagnosis, the collection of subjective assessments of medical images concerning aesthetic aspects is also a possible use-case. The study of Vartanian et al. [18] focuses on the definition of the ideal thigh proportions. Therefore, the authors analyze the ratings of crowdworkers concerning the perceived attractiveness of thighs shown on photographs. While this work only discusses mostly medical aspects, the work of Tse et al. [16] also consider the reliability of crowdsourced ratings and compare them to assessments provided by experts. Here, the aim of the study is the evaluation of the aesthetic outcome of treatments for unilateral cleft lip. The results of the study show that the ratings of crowdworkers are reliable and well correlated to the assessments of the expert group.

Even if the study of Tse et al. discusses and compares assessments of crowdworkers and an expert group, they did not consider the group of affected persons and their friends or family members. Furthermore, there is only a small amount of research about the comparability of crowd-based assessments and expert ratings in the context of aesthetic, medical cases. Thus, it is unclear if these results are transferable to other medical fields. In this work, we do not only discuss assessments of crowdworkers and specialists, but we also consider the perception from affected people, i.e., parents of children with deformational cranial asymmetries, as well as from a non-crowdsourcing layperson group and a group of physicians who are no pediatricians.

3 Study Description

In this section, we describe the design of the user study. Additionally, we detail on the preprocessing of the medical data as well as the conduction of the study.

3.1 Medical Dataset

Our dataset consists of 3D scans from 51 newborns' heads that exhibit different severities and types of deformational cranial asymmetries. In addition, different objective asymmetry metrics are available for each patient, e.g., the Cephalic Index. While it is desirable to collect subjective rating for the whole dataset, this is not reasonable for this preliminary test. To perform a comparison of the ratings of the different groups of test-takers, a larger number of ratings per stimulus and group is necessary. While affected parents might be tolerant against a long-lasting subjective test that includes all scans, practitioners that participate voluntarily might not be willing to spend too much time on this research task. Further, paid crowdsourcing tasks should also be kept short, as long tasks fatigue workers and workers might start to rush through the task instead of performing it thoroughly. Another option would be to split the dataset into smaller subsets. However, as the number of practitioners and parents are

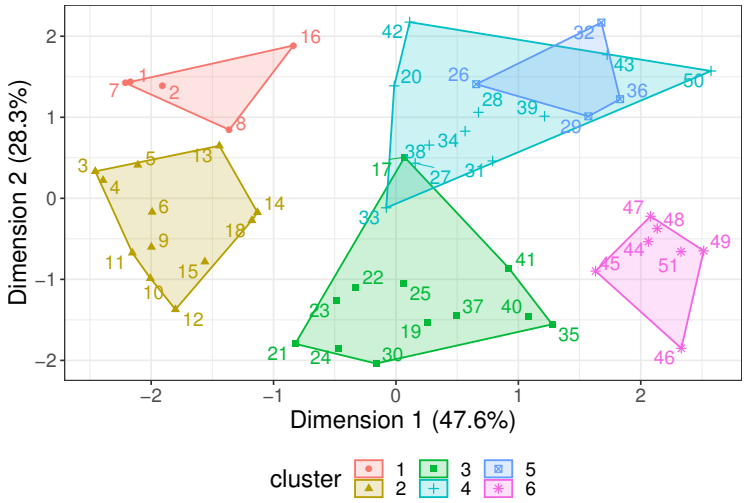

Fig. 1. Clustering of the 3D scans based on the patients medical data.

highly limited in our case, we decided to focus on a larger number of ratings per scan, instead of a large number of annotated scans.

We use clustering to minimize the number of annotated samples but still test a diverse and representative set. In our specific case, we decided to cluster the patients with the Partitioning Around Medoids (PAM) algorithm, as the PAM algorithm identifies data points as cluster centers instead of calculating theoretical cluster centers like, e.g., k-means [4]. As the distance metric, the Euclidean distance is calculated. The objective medical metrics of the participants' heads, described in Section 2, are normalized to have zero mean and unit variance and are used as features for the clustering. An elbow plot is used to identify a suitable number of clusters for initializing the process. Further, the clustering is evaluated by using silhouette coefficients, and based on the results, the optimal number of clusters is identified as six. Figure 1 shows a two dimensional representation of the final clustering.

Based on the clustering, we select a representative patient for each of the six clusters. These representatives are as distinct as possible concerning their characteristics. Four additional patients are added based on the suggestion of medical experts. This results in a dataset of 10 different patients for the evaluation. To further evaluate if an asymmetry on the left or right side of the head is perceived differently, we also generated mirrored versions of all scans.

3.2 Study Design

The original 3D scans are produced with the methodology described by Meyer-Marcotty et al. [9] that allows viewing the scan from different perspectives interactively. However, in order to make the scans accessible to a large group of persons, we have to guarantee that they can be viewed on all types of devices

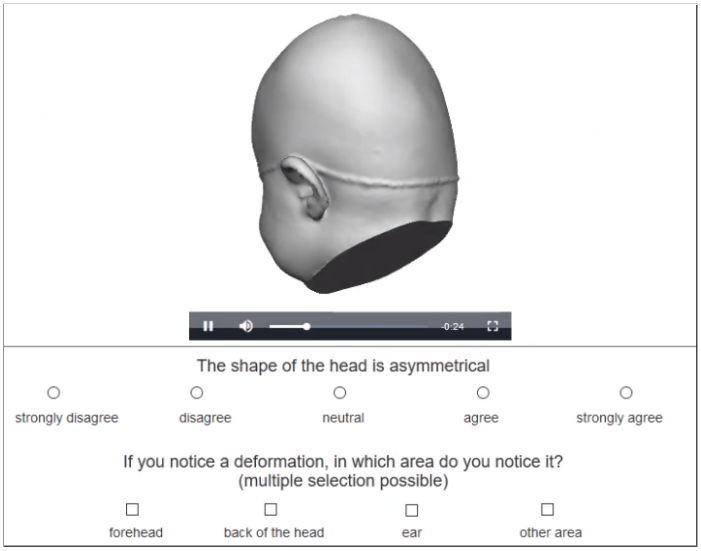

Fig. 2. Mockup of the web page showing a video of the rotating head.

with a minimum amount of user preparations. Therefore, we decide to convert the scans to short videos of 30 seconds. Each video shows a rotating head omitting the frontal view of the newborn's face due to privacy policies. Below the video, the test takers are asked to judge if the head is asymmetrical on a five-point absolute category rating scale. In case the test taker recognizes a deformation, the test taker is also asked to indicate the area of the head where the asymmetry has been noticed, e.g., at the back of the head or the forehead.

To make the study accessible for experts, parents, and laypersons, we realize it as a web page containing the following steps. After introducing the participants to the subject of the study, the videos of the scanned heads were shown to the test taker, namely the original video of the scans of the ten selected patients and the mirrored version of these videos. Two of the videos are shown twice to evaluate the constancy of the ratings. In total, each participant watches 22 videos in random order. Figure 2 shows a screenshot of the realization of the web page containing a video.

After rating all videos, the participants are asked to provide additional demographical information, like age and gender. Further, we collect information if the participant works in health care, if so in which area as well as if the participant has children in the age between zero to six years. At the end of the survey, the participant has the option to give additional feedback.

3.3 Study Conduction

The group of crowdworkers has been recruited via the crowdsourcing platform Microworkers[6] in March and April 2017. We limited the study to users from the

[6] `https://microworkers.com` Accessed Jun. 2019

Table 1. Overview about age and gender of the groups of participant.

Group	N	∅ Age	Female [%]
$Pediatricians$	31	50	25.1%
$OtherPhysicians$	27	42	48.1%
$Parents$	73	38	76.7%
$Crowdworkers$	54	32	46.3%
$OtherNon-Experts$	54	42	61.1%

United States, Canada, and the United Kingdom to prevent misunderstandings concerning the instructions due to language barriers. Further, the limitation reduces side effects due to demographic or cultural differences, e.g., aesthetic aspects. The payment per participation was $0.50. The participants of the other groups, i.e., pediatricians, physicians, and parents, as well as the other non-experts, have been invited via e-mail between July and September 2017. Here, participation has been voluntary.

Table 1 presents an overview of the groups of participants. We only consider participants who answer all questions. Overall, 54 crowdworkers take part in our study. Those workers are on average 32 years old, and 46.3% of the group is female. The average age of the parents (38 years), other non-experts (42 years) and other physicians (42 years) is slightly higher than in the group of crowdworkers. Further, the group of pediatricians is the oldest (50 years) with the lowest share of female participants (25.1%), while the group of parents has the highest share of female members (76.7%).

4 Results

In this section, the constancy of the provided ratings is analyzed. Further, we compare the ratings of the groups of participants and discuss factors which may influence the perception of the groups.

4.1 Constancy of Ratings

To evaluate the constancy of the answers given by the participants, we compare the ratings of the videos which are shown twice within the study, i.e., the video of patient 1 and patient 17. We analyze if the ratings of the first and the second occurrence of the videos originate from the same distribution by using the Kruskal-Wallis rank sum test. The test results in a rejection of the null hypothesis ($\chi^2 = 28.83, df = 4, p < 0.001$). A pairwise comparison of the samples per group using the Wilcoxon rank sum test with Bonferroni correction shows significant differences between the crowdworkers and the other non-experts ($p < 0.01$), parents ($p < 0.001$), pediatricians ($p < 0.01$) and other physicians ($p < 0.01$). We found no significant differences between the other groups ($p > 0.05$). By comparing the mean, standard deviation and quantiles of the answers of each group, a higher divergence between the ratings provided by the crowdworkers and the other groups are seen (see Table 2).

Table 2. Statistical parameters of rating differences of videos shown twice during the study.

Group	Mean	SD	90% Quantile
$Pediatricians$	0.29	0.49	1
$OtherPhysicians$	0.25	0.44	1
$Parents$	0.32	0.52	1
$Crowdworkers$	0.80	0.93	2
$OtherNon-Experts$	0.32	0.51	1

Especially, the 90% quantile indicates that the crowdworkers' ratings for watching a video the second time often differ for more than one point on the rating scale. This effect may be explainable by a training phase that is more noticeable for the crowdworkers. If so, the participants should be more precise in rating the asymmetry of later shown videos especially for the copies, i.e., by answering with the option *(strongly) agree/disagree* instead of selecting the neutral one. Therefore, the correlation between the absolute position of the copied videos and the distance of the selected options to the neutral option is analyzed. Other than expected, we found no significant relationship between these values by using Spearman's rank correlation. This result indicates that on the one hand, there are other crowd-specific factors which may lead to different ratings for some participants. On the other hand, it may be an indicator that these participants are inattentive.

In the following evaluations, we only consider participants who submit constant ratings. This means we exclude participants providing ratings of the videos shown twice, which are not identically or are not located next to each other in the rating scale. Overall, 21 crowdworkers, two other non-experts, four parents, and one pediatrician are filtered out. Further, the assessments of the copied videos are omitted from the evaluation.

4.2 Comparison of Asymmetry Ratings

To evaluate potential effects on the perceived asymmetry caused by the background, e.g., previous knowledge about deformational cranial asymmetries, we analyze the assessments of the participants per group and compare the mean opinion. A rating of 1 represents the option *strongly disagree*, the value 3 corresponds to a neutral rating while a rating of 5 means that the participant *strongly agrees* that the shown head is asymmetrical. In the following evaluation, we omit the mirrored version of the videos to prevent biases caused by side effects.

To analyze differences in the ratings between the groups, we run a one-way ANOVA. The test shows a significant effect of the group on the ratings ($F(4, 2075) = 93.81$, $p < 0.001$). Bonferroni's post-hoc test revealed significant differences between the crowdworkers and all other groups ($p < 0.001$). By analyzing the ratings in detail, we observe that 62.2% of the crowd-based ratings (strongly) agreed that the shown heads are deformed. In comparison to the other groups with a percentage of agreements ranging from 7.7% to 14.2%, the amount

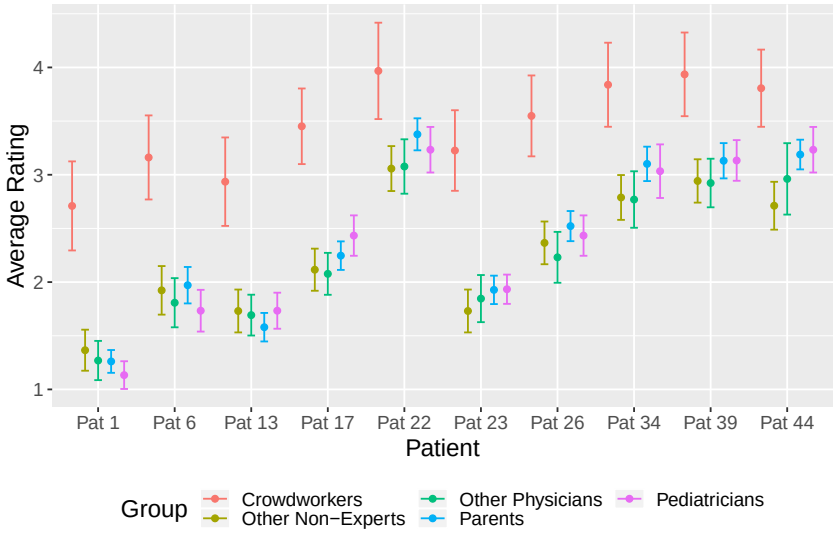

Fig. 3. Mean of the ratings with 95% confidence intervals per group.

is by far higher. Other than expected, this indicates that the crowdworkers perceive weak deformations as more critical than the other groups. An explanation may be crowd-specific, additional influence factors on the assessments, e.g., the participants are less attentive due to distractions or rushing through the test. Alternatively, the phrasing of our question might induce bias, and the workers might assume that they are *expected* to identify an asymmetry.

As the observation may be invalid on a per patient basis, the assessments are also evaluated on a patient level. The average ratings per patient, including the 95% confidence intervals, are shown in Figure 3. While the average ratings for the groups of other non-experts, parents, pediatricians, and other physicians are quite similar with mostly overlapping confidence intervals, the crowdworkers rate more often neutral or agreed to notice a deformation. This leads to higher average ratings with a constant offset of approximately one point on the rating scale for each patient. This observation corresponds to the result of a one-way ANOVA per patient, which shows a significant difference between the ratings of the groups for all patients ($p < 0.001$). By using the Bonferroni post-hoc test again, a significant difference between the assessments of the crowdworkers and those of all other groups is revealed for all patients ($p < 0.05$). Between the other groups, we found no significant differences except for *patient 44*. For this patient, there is a significant difference between the group of other non-experts and the parents ($p < 0.01$) as well as other non-experts and pediatricians ($p < 0.05$).

The different findings for the patients indicate that for unique characteristics of deformation the perception differs between experts (physicians and parents) and laypersons. We further analyze this aspect by evaluating the assessments of the mirrored and original videos as well as the provided answers concerning the areas where the participants noticed the deformations.

Table 3. Coefficients r of point-biserial correlation between ratings and areas of noticed deformation, i.e. front head, back of the head, ear and other areas including level of significance.

Group	r Front	r Back	r Ears	r Other
$Pediatricians$	0.47***	0.64***	0.48***	0.20***
$OtherPhysicians$	0.27***	0.57***	0.38***	0.19**
$Parents$	0.33***	0.64***	0.39***	0.11**
$Crowdworkers$	0.07	0.42***	0.30***	0.11
$OtherNon-Experts$	0.17***	0.49***	0.27***	0.26***

** 0.01, *** 0.001

4.3 Influence Factors on the Perceived Asymmetry

As it may influence the perception if the deformation is located on a head's left or right side, the assessments of the original and the mirrored scans are evaluated. By using a repeated-measures ANOVA, no significant differences between the ratings of the original and the mirrored videos for all groups could be found ($p > 0.05$). Thus, perception is not influenced by this aspect.

The relation between the ratings and the answers to the question in which area the deformation has been noticed is analyzed, to get a better understanding of the test takers' ratings. Table 3 summarizes the correlation coefficients per group between the ratings and the given answers. For all groups, a significant, positive correlation between the ratings and the selection of the back of the head and the ears as the noticed location of deformation is observed. The crowdsourced assessments do not significantly correlate with their answers concerning the forehead and the selected option *other areas*, while for these options a significant, positive correlation for the other groups is seen. Correlations between noticed deformations at the front head and the rating are higher for the groups of pediatricians, other physicians and parents. This indicates on the one hand that for non-experts it is more challenging to identify deformations at the front of the head due to the missing view of the face of the newborns. On the other hand, it may be a piece of evidence that the groups focus on different areas of the head, which may influence the perception.

5 Conclusion

Utilizing crowdsourcing for the collection of subjective assessments, e.g., for evaluating the perceived quality of technical systems by the users, is a commonly used approach in several research directions. Nowadays, collecting ratings for medical use cases via crowdsourcing gain more and more interest.

In this work, we introduce a novel medical use-case for the collection of ratings about the perceived severity of head deformations. We conducted a user study involving people with different background, i.e., pediatricians, other physicians, parents of children with deformational cranial asymmetries and non-experts including crowdworkers and non-crowdworkers, leading to different prior knowledge about deformational cranial asymmetries.

The results of the study showed that the perception of the crowdworkers and the other groups differ when comparing the ratings independent from the patients. While the crowdworkers more often perceived deformation of the shown heads, the other non-expert group rates mostly similar to the groups of physicians and parents. Here, the similar perception of experts, affected people, and laypersons is other than expected.

The analysis of the ratings per patient showed that some characteristics of deformations also leads to differences in the perception of the laypersons and the expert groups. Further, we found that the recognition of deformations is based on different areas of the head for people with a medical background and affected persons. Even if a frontal view of the faces is not shown, which makes it challenging to notice deformations on the forehead, they consider this area for their ratings. This observation may be an explanation of the different perception of deformational cranial asymmetries, as mentioned above.

Nevertheless, the differences concerning the focus of the participants do not fully explain the variations in the ratings of the crowdworkers and the other groups. Instead, these ratings may be influenced by crowd-specific, additional factors, e.g., inattentiveness due to distractions, biases induced by the phrasing of the instructions, or an insufficient training phase, which will be subject of future research.

Furthermore, the objective medical information could be considered to get a more in-depth insight into the relationship between the objective and the subjective data.

Overall, the results of our study encourage the involvement of a diverse group of people with different knowledge and background concerning the subject of studies. Further, our observations show the importance of carefully designing such studies when conducting them in the context of crowdsourcing.

Acknowledgment

The authors thank Veronika Cheplygina for the fruitful discussions on crowdsourcing in the context of medical images and Norman Stulier for his support during the implementation of the survey software. This work is supported by Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2-2, TR 257/38-2. The authors alone are responsible for the content.

References

1. Alialy, R., Tavakkol, S., Tavakkol, E., Ghorbani-Aghbologhi, A., Ghaffarieh, A., Kim, S.H., Shahabi, C.: A review on the applications of crowdsourcing in human pathology. Journal of pathology informatics **9** (2018)
2. Argenta, L.: Clinical classification of positional plagiocephaly. Journal of Craniofacial Surgery **15**(3) (2004)
3. Captier, G., Dessauge, D., Picot, M.C., Bigorre, M., Gossard, C., El Ammar, J., Leboucq, N.: Classification and pathogenic models of unintentional postural cranial

deformities in infants: plagiocephalies and brachycephalies. Journal of Craniofacial Surgery **22**(1) (2011)
4. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
5. van der Heijden, L., Piner, S.R., van de Sande, M.A.J.: Pigmented villonodular synovitis: a crowdsourcing study of two hundred and seventy two patients. International orthopaedics **40**(12) (2016)
6. Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. IEEE Transactions on Multimedia **16**(2) (2014)
7. Li, Y., Du, N., Liu, C., Xie, Y., Fan, W., Li, Q., Gao, J., Sun, H.: Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers from non-experts. In: Proc. of the International Conference on Web Search and Data Mining (2017)
8. Meyer, A.N., Longhurst, C.A., Singh, H.: Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of crowdmed. Journal of medical Internet research **18**(1) (2016)
9. Meyer-Marcotty, P., Boehm, H., Linz, C., Kunz, F., Keil, N., Stellzig-Eisenhauer, A., Schweitzer, T.: Head orthesis therapy in infants with unilateral positional plagiocephaly: an interdisciplinary approach to broadening the range of orthodontic treatment. Journal of Orofacial Orthopedics **73**(2) (2012)
10. Ørting, S., Doyle, A., Hirth, M., van Hilten, A., Inel, O., Madan, C.R., Mavridis, P., Spiers, H., Cheplygina, V.: A survey of crowdsourcing in medical image analysis. arXiv preprint arXiv:1902.09159 (2019)
11. Peleg, M., Leung, T.I., Desai, M., Dumontier, M.: Is crowdsourcing patient-reported outcomes the future of evidence-based medicine? a case study of back pain. In: Proc. of the Conference on Artificial Intelligence in Medicine in Europe (2017)
12. Persing, J., James, H., Swanson, J., Kattwinkel, J., on Practice, C., Medicine, A., et al.: Prevention and management of positional skull deformities in infants. Pediatrics **112**(1), 199–202 (2003)
13. Ranard, B.L., Ha, Y.P., Meisel, Z.F., Asch, D.A., Hill, S.S., Becker, L.B., Seymour, A.K., Merchant, R.M.: Crowdsourcingharnessing the masses to advance health and medicine, a systematic review. Journal of general internal medicine **29**(1) (2014)
14. Salloum, W., Edwards, E., Ghaffarzadegan, S., Suendermann-Oeft, D., Miller, M.: Crowdsourced continuous improvement of medical speech recognition. In: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence (2017)
15. Servadei, L., Schmidt, R., Eidelloth, C., Maier, A.: Medical monkeys: A crowdsourcing approach to medical big data. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer (2017)
16. Tse, R.W., Oh, E., Gruss, J.S., Hopper, R.A., Birgfeld, C.B.: Crowdsourcing as a novel method to evaluate aesthetic outcomes of treatment for unilateral cleft lip. Plastic and reconstructive surgery **138**(4) (2016)
17. Tucker, J., Day, S., Tang, W., Bayus, B.: Crowdsourcing in medical research: theory and practice. Tech. rep., PeerJ Preprints (2018)
18. Vartanian, E., Gould, D.J., Hammoudeh, Z.S., Azadgoli, B., Stevens, W.G., Macias, L.H.: The ideal thigh: a crowdsourcing-based assessment of ideal thigh aesthetic and implications for gluteal fat grafting. Aesthetic surgery journal **38**(8) (2018)
19. Wilbrand, J.F.: Transferring the assessment of cranial deformities to the affected. Journal of Craniofacial Surgery **28**(2) (2017)