MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220130

User-Centered Evaluation of a Visual Annotation Tool for Rapid Assessment of Pediatric Weight Entry Errors

Danny T.Y. Wu, PhD, MSI[a,b], PJ Van Camp, MD[a,d], Abraham Kim[a,c], Milan Parikh[a,c], Lei Liu[a,d], Monifa Mahdi[d], Yizhao Ni, PhD[d,b], S. Andrew Spooner, MD, MS[d,b]

[a]*Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, OH, USA*
[b]*Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA*
[c]*Medical Sciences Baccalaureate Program, University of Cincinnati College of Medicine, Cincinnati, OH, USA*
[d]*Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*

Abstract

Weight entry errors can cause significant patient harm in pediatrics due to pervasive weight-based dosing practices. While computerized algorithms can assist in error detection, they have not achieved high sensitivity and specificity to be further developed as a clinical decision support tool. To train an advanced algorithm, expert-annotated weight errors are essential but difficult to collect. In this study, we developed a visual annotation tool to gather large amounts of expertly annotated pediatric weight charts and conducted a formal user-centered evaluation. Key features of the tool included configurable grid sizes and annotation styles. The user feedback was collected through a structured survey and user clicks on the interface. The results showthat the visual annotation tool has high usability (average SUS=86.4). Different combinations of the key features, however, did not significantly improve the annotation efficiency and duration. We have used this tool to collect expert annotations for algorithm development and benchmarking.

Keywords:

Usability Testing; Data Visualization; Data Curation

Introduction

Research has shown that pediatric weight error can happen more than 150 times per year in electronic health records (EHRs) [1]. These weight errors entered by health professionals can cause significant harm to patients, especially in pediatrics where weight-based dosing is widely used. To address this issue, computational methods have been developed to detect anomalies in anthropometric data to improve care quality and clinical research, including methods from the Centers of Disease Control (CDC) [2, 3], Children's' Hospital of Philadelphia (CHOP) [4], and Cincinnati Children's Hospital Medical Center (CCHMC) [5]. These methods, however, do not have outstanding sensitivity and accuracy and therefore cannot be implemented as an effective clinical decision support system (CDSS) to be used in care routines. Moreover, these methods cannot fully capture abnormal weight values with high clinical significance, i.e., whether an abnormal weight value is worth clinicians' attention during busy clinics to prevent medication errors [6]. Therefore, it is critical to develop a novel method with machine intelligence to identify abnormal weight errors for research and quality improvement purposes and in both retrospective and prospective manners.

In the past few years, we have been working as an interdisciplinary team to investigate this issue. This team consists of informatics experts from hospital medicine, system evaluation, and machine learning (ML). Our ultimate goal is to develop and disseminate a highly usable computational method to identify abnormal weight values and deploy it as a CDSS. In our previous research [6], we have found that medically trained experts (e.g. attending physicians, fellows, and residents) can review growth charts and quickly spot abnormal weight values with high clinical importance. Based on this finding, we framed the algorithm development as a supervised learning process, which requires collection of expert annotation to train the ML algorithms.

We therefore designed an innovative annotation tool that utilizes the seven visual analytics design principles [7]. This visual annotation tool (VAT) was developed and preliminarily evaluated for its potential in rapidly collecting large-scale expert-annotated weight errors, which was published in MedInfo 2019 [8]. In this paper, we reported the results of the formal user-centered evaluation of the VAT as the final step to conclude the tool development. It is worth noting that the evaluation was focused on the usability of the VAT, rather than the actual collection of expert annotations.

Methods

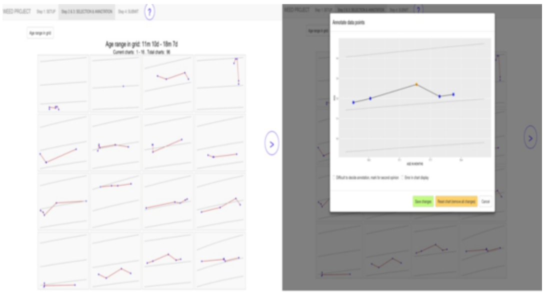
Visual Annotation Tool (VAT)

The VAT was developed as a web-based application using R Studio1 and its Shiny library2. The annotation datasets were stored in an open-sourced SQLite database3. The development of the VAT followed agile software development principles with three phases to refine the tool [8]. The VAT has two key features to facilitate the growth chart review process: grid size (configurable number of growth charts displayed on the screen at once) and annotation styles (annotation process). Annotation style refers to either one-step or two-step. One-step annotation means that a growth chart is annotated immediately when it is selected from the grid view. Two- step annotation saves all the selected charts in an initial screening phase and then displays them for annotation in the second phase. An example of the

[1] https://rstudio.com/
[2] https://shiny.rstudio.com/
[3] https://www.sqlite.org/

One-step annotation isshown in Figure 1. On the left-hand side of the figure, users were able to screen the growth charts in a grid and select one chart for detailed review. On the right-hand side of the figure, users can further select a data point onthis enlarged chart, should they believe it was an error. All data points were defaulted to blue (not an error). By clicking each point, the color will be changed from blue to orange (likely an error) and red (definitely an error) then back to blue. The source code of the VAT can be retrieved on GitHub[4]. Of note, the current version of the VAT was not connected to an electronic health record system. The weight data were extracted, de-identified, and sampled in a separate process.

Figure 1 – User interface of the visual annotation tool,annotation page. Note on the left is a screen with an overall view of the 16 charts and on the right is screenthat displays when pressing on one of the charts.



Participant Recruitment

The participants were recruited on a voluntary basis using convenience sampling through the professional network of the research team at the CCHMC and UC. All participants had medical training and knowledge and indicated that they had adequate familiarity with interpreting pediatric weight charts prior to recruitment. The participants consisted of three pediatric attendings, seven pediatric residents and fellows, and six medical students who had completed their required pediatric rotation. More than half of the participants were in the age range 26-35 (n=10, 62.5%), followed by a quarter in the age range 18-25 (n=4, 25%). A quarter of the participants were females (n=4, 25%).

Study Conduction

This study conducted a lab experiment to evaluate the usability of the VAT. The participants were assigned todifferent settings as shown in Table 1. The study design included the use of a structured survey, interviews, and analysis of user clicks. The participants were invited to aone-hour session to test the usability and functionality ofthe VAT. During the session, the participant was introduced to the VAT and annotated two datasets (denoted as Dataset 1 and Dataset 2), each consisted of 225 growth charts. The participant then filled out a survey (see data collection for details). Each participant received a $20 reloadable debit card for compensation. Of note, the patient data were de-identified to remove protected health information following the standard of the Health Insurance Portability and Accountability Act (HIPAA) since the purpose of the present study was to evaluate the usability of the VAT, not collecting expert- annotations for the actual ML model training. The study protocol was reviewed and approved by the internal review board at UC (#2017-2075).

Table 1 – Participant Assignment and Grouping

Participant	Controlled Variable	Experimental Variable	Dataset 1 Setting	Dataset 2 Setting
P01 P09	One-step annotation	Grid Size	Grid Size:9	Grid Size:25
P02 P10	Two-step annotation		Grid Size: 9	Grid Size:25
P03 P11	One-step annotation		Grid Size:25	Grid Size:9
P04 P12	Two-step annotation		Grid Size:25	Grid Size:9
P05 P13	Grid Size: 9	Style of Annotation	One-step	Two-step
P06 P14	Grid Size: 25		One-step	Two-step
P07 P15	Grid Size: 9		Two-step	One-step
P08 P16	Grid Size: 25		Two-step	One-step

Data Collection

To collect user feedback on this tool, a 16-question survey in combination with analysis of the click data were analyzed. The structured survey included questions from the System Usability Scale (SUS) [9] (N=10) and additional custom questions (N=6) on a 5-point Likert scale. Table 2 lists all of the survey questions with a sample score for showing the calculation of a SUS composite score. User clicks were captured by the VAT in the form of event logs. The data consisted of a list of actions related to clicks in various parts of the application and an associated timestamp. All timestamps were encoded in the Unix timestamp format. Each action, predefined by the research team, was ordered consecutively with an identifier associated with the type of action. The action descriptions included which graphs were selected, which data points were annotated and changed, whether charts were zoomed in or out, and which buttons were clicked.

Data Analysis

The first part of the analysis focused on survey data. The first ten SUS questions were scored using the original mechanism and generated a composite score for each participant [9]. Specifically, the composite score was calculated by subtracting one from the answers of the odd numbered questions (i.e. answers for QID 1, 3, 5, 7, and 9), subtracting the answers of the even numbered questions from five (i.e. answers for QID 2, 4, 6, 8, and 10), and multiplying the sum of all these values by 2.5. An example calculation is provided using the scores in Table

3. Each participant received a composite score based on their answers. The average composite score of all 16 participants was calculated and then compared to a threshold value (68) to indicate the usability level of the system. In other words, the VAT is considered as above average usability if the average composite score is higher than 68 [9]. The six additional questions in the survey were summarized statistically through the minimum, maximum, median, average, and standard deviation of the scores of each question.

The second part of the analysis involved the event log dataand focused on the efficiency of the VAT to support chartreview and annotation. Two measures were used: click and duration. Click

[4] https://github.com/iCDCU-lab/open_VATCHE

Table 2 – Survey evaluating the usability of the VAT.

ID	Question	SUS*
1	I think that I would like to use this system frequently.	3
2	I found the system unnecessarily complex.	1
3	I thought the system was easy to use	5
4	I think that I would need the support of a technical person to be able to use thissystem.	1
5	I found the various functions in this system were well integrated.	4
6	I thought there was too much inconsistency in this system.	1
7	I would imagine that most people would learn to use this system very quickly.	4
8	I found the system very cumbersome to use.	1
9	I felt very confident using the system.	4
10	I needed to learn a lot of things before I could get going with this system.	2
11	The use of a grid helped me to quickly find only charts that were of interest to me.	
12	The display of the weight charts was intuitive considering my background on interpreting traditional weight charts.	
13	The app allowed me all manipulations (viewing chart in detail, zooming, getting information on a single point) I needed for good analysis.	
14	The color scheme was intuitive (blue, red, and orange)	
15	There was no significant lag or loading time that slowed down the workflow.	
16	I encountered no technical errors.	

* Sample values for calculating the SUS composite score=((3-1)+(5-1)+(4-1)+(4-1)+(4-1)+(5-1)+(5-1)+(5- 1)+(5-1)+(5-2))*2.5=85

was defined as the total number of clicks required for a participant to finish each dataset annotation. Once records per participant per dataset were identified, the duration of a dataset was calculated. Duration was defined as the difference between the timestamps at the start and end record per participant per dataset. Fifteen comparison groups were generated in Table 3, comparing grid sizes, annotation styles, and datasets. Since each group was tested on both efficiency and duration measures, a total of 30 comparisons were conducted. Further combinations were not feasible since the records per group (sample size) would be small (N=4). The means of the scores were compared using two-tailed t-test with nonequal variance.

Results

Table 4 lists the composite SUS score for each participant. Overall, the participants scored high on the VAT usability, with composite scores ranging from 77.5 to 97.5. The average score was 86.4, indicating the above- average us- ability of the VAT.

Results for the additional six questions are summarized in Table 5. While the overall usability is high, the additional questions suggest areas for improvement. Most participants rated the VAT highly on the use of grid size and the intuitive display of weight charts (QID 11 and 12) with high consensus (smaller standard deviation). The participants also indicated good functionality of the VAT to support the annotation process (QID13 and 14). However, some participants reported lag or technical errors within the tool (QID 15 and 16). These issues have been noted for the next iteration of tool development.

Table 3 – Compared Groups and Controlled Variables

Group ID	Controlled variable	Comparison	Participants in the group
1	None	Grid Size 9 vs. 25	16
2		One vs. two-step	16
3		Dataset 1 vs. 2	16
4	One-step	Grid Size 9 vs. 25	8
5		Dataset 1 vs. 2	8
6	Two-step	Grid Size 9 vs. 25	8
7		Dataset 1 vs. 2	8
8	Grid Size 9	One vs. two-step	8
9		Dataset 1 vs. 2	8
10	Grid Size 25	One vs. two-step	8
11		Dataset 1 vs. 2	8
12	Dataset 1	Grid Size 9 vs. 25	8
13		One vs. two-step	8
14	Dataset 2	Grid Size 9 vs. 25	8
15		One vs. two-step	8

For user clicks, a total of 3,533 records were collected from the 16 participants. On average, each dataset took 104 clicks and 572 seconds (less than 10 minutes) to finish. All of the comparisons in Table 3 on the click and the duration measure were not statistically significant, except the #11 comparison of duration between Dataset 1 and Dataset 2 when the grid size 25 was controlled. That is, when the grid size is larger, the participants annotated the same growth charts faster with a 45% decrease (from 742 to 410 seconds, p=0.0345). It seems that a larger grid size can better facilitate the annotation process. Although this difference may be contributed by the learning effect of the participants, the comparisons between datasets in other controlled variables (one-step, two-step, grid size 9) did not show any significance. In other words, if learning effect was a dominant factor, all comparisons between datasets should have been significant. These results indicate that there were multiple factors affecting the annotation efficiency, and none of these were dominant.

Discussion

This user-centered evaluation showed that the VAT can achieve a high usability, with an average SUS score of 86.4, substantially higher than the threshold of 68. The VAT was able to support efficient annotation, with an average duration of 10 minutes on 225 growth charts. However, different combinations of the key features of the VAT, i.e. grid size and annotation style, did not significantly improve annotation efficiency. The only exception was the significant decrease in time taken in Dataset 1 to Dataset 2 with the grid size of 25. Since the same comparison was not significant under the grid size of 9, a larger grid size seemed to improve the annotation efficiency and promote learning effect. The fact that mostofthe comparisons were not significant suggests that the VAT should provide flexible configurations of the key features.

Table 4 – Composite SUS score for each participant.

The average SUS Composite Score: 86.4 ± 6.8

Participant	Composite Score	Participant	Composite Score
P001	77.5	P009	92.5
P002	85.0	P010	85.0
P003	80.0	P011	97.5
P004	80.0	P012	97.5
P005	85.0	P013	87.5
P006	80.0	P014	92.5
P007	95.0	P015	87.5
P008	77.5	P016	82.5

Researchers who would like to develop their own informatics tools to support user annotation and other research tasks should consider the following key takeaways of our work: 1) apply visual analytics principals to better support labor-intensive tasks, 2) consider clinicians' busy schedule and use multiple means to collect user feedback, and 3) conduct user- centered evaluation in phases with experts and pseudo/real users and iteratively refine the tool. Additionally, researchers may extend the design of the VAT, such as changing a square grid to a rectangle gird to better fit a wide computer screen.

Table 5 – Summary statistics of the second part of the survey.

ID	Short Description	Min	Median	Max	Average (SD*)
11	Grid can help.	4	5	5	4.75 (0.45)
12	Intuitive display.	4	5	5	4.81 (0.40)
13	Functions to support.	2	4.5	5	4.25 (1.00)
14	Intuitive color-coding scheme.	3	5	5	4.38 (0.89)
15	No significant lag time.	1	3	5	3.19 (1.05)
16	No technical errors.	1	5	5	4.19 (1.22)

* SD = standard deviation

We have started using the VAT to collect expert annotations for algorithm development. We have recruited 18 medical experts to annotate a sample of15,000 pediatric growth charts using stratified sampling. Since we preferred each growth chart to be reviewed by three experts, each expert has received 2,500 growthcharts for annotation. The individual dataset has been split into 10 subsets (N=250) to better fit into the experts' busy clinical schedule. Based on the experimental results of the current study, we have estimated that each expert is expected to spend 100-120 minutes (or 2 hours) to finish the task. Multiple algorithms including support vector machines, random forests, and artificial neural networks will be trained and tested on the annotated datasets. The results of the algorithm development will be reported in future publications.

Limitations

This evaluation study has several limitations. First, we did not use methods such as think-aloud protocol [10], card sorting [11], qualitative interviews, or screen recording, to collect more user feedback. However, we believe the current two methods, namely structured survey and analysis of event logs, successfully collected enough user feedback to determine the usability and efficiency of the VAT and identify opportunities for improvement. Second, the VAT was not evaluated on the accuracy of their annotations, nor was it evaluated in a prospective manner (i.e., presenting an error without knowing what has happened afterwards) since the focus of the user- center evaluation was on the usability and efficiency of the VAT. Moreover, the annotation datasets were served as a reference standard, not a gold standard since no chart reviews were done to determine the actual weight entry errors.

In the third place, the comparison of the key features was limited to one controlled variable due to the number of participants. Similarly, the study is limited in statistical power due to the lack of resources to recruit more participants. This limitation can be addressed in the future work to include a representative sample of the whole user population. Currently, the VAT allows users to choose their preferred configurations to conduct annotation tasks. Next, the event logs were analyzed at the macro level (clicks and durations), not at the micro level (e.g., sequential patterns of the clicks). Although the latter could have generated interesting patterns to explain the user behaviors, the current analysis is sufficient to explore any dominant factor affecting the annotation efficiency of the VAT. Last but not least, the VAT did not provide any clinical context of the patient, such as diagnosis and medication information, because our goal of the ML development was to determine the abnormality of a weight point solely based on the trend in a growth chart.

Conclusions

We developed and evaluated the VAT in a user-centered manner to demonstrate its high usability and efficiency. We have started to use the VAT to collect a large amountof labeled data for algorithm development. The ML algorithms will be further developed as an intelligent CDSS to be used prospectively in clinical work andretrospectively on legacy data to improve clinical data quality and patient safety.

Acknowledgements

This study was partially supported by the Mary K. Logan Award received by Dr. Danny T.Y. Wu in 2018. The award was provided by the Association for the Advancement of Medical Instrumentation. We thank Ms. Anunita Nattam for her effort on proofreading the manuscript.

References

[1] Hagedorn PA, Kirkendall ES, Kouril M, *et al.* Assessing Frequency and Risk of Weight Entry Errorsin Pediatrics. *JAMA Pediatr* 2017;171:392–3. doi:10.1001/jamapediatrics.2016.3865

[2] Kuczmarski RJ, Ogden CL, Guo SS, *et al.* 2000 CDC Growth Charts for the United States: methods and development. *Vital Health Stat 11* 2002;:1–190.

[3] A SAS Program for the 2000 CDC Growth Charts (ages 0 to <20 years).

https://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm (accessed 6 Mar 2018).

[4] Daymont C, Ross ME, Russell Localio A, *et al.* Au-tomated identification of implausible values in growth data from pediatric electronic health records.*J Am Med Inform Assoc JAMIA* 2017;24:1080–7. doi:10.1093/jamia/ocx037

[5] Spooner S, Shields S, Dexheimer J, *et al.* Weight En-try Error Detection: A Web Service for Real-time Statistical Analysis. 2016.

[6] Wu DTY, Meganathan K, Newcomb M, *et al.* A Comparison of Existing Methods to Detect Weight Data Errors in a Pediatric Academic Medical Center.*AMIA Annu Symp Proc AMIA Symp* 2018;2018:1103–9.

[7] Conlen M, Stalla S, Jin C, *et al.* Towards Design Principles for Visual Analytics in Operations Contexts. In: *Proceedings of the 2018 CHI Conference onHuman Factors in Computing Systems - CHI '18*. Montreal QC, Canada: : ACM Press 2018. 1–7. doi:10.1145/3173574.3173712

[8] Van Camp PJ, Mahdi CM, Liu L, *et al.* Developmentand Preliminary Evaluation of a Visual Annotation Tool to Rapidly Collect Expert-Annotated Weight Er-rors in Pediatric Growth Charts. *Stud Health Technol Inform* 2019;264:853–7. doi:10.3233/SHTI190344

[9] Brooke J. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Read UK Digit Equip Co Ltd* 1986;43.

[10] Alhadreti O, Mayhew P. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Proto- cols. In: *Proceedings of the 2018 CHI Conference onHuman Factors in Computing Systems - CHI '18*. Montreal QC, Canada: : ACM Press 2018. 1–12. doi:10.1145/3173574.3173618

[11] Wood JR, Wood LE. Card Sorting: Current Prac-tices and Beyond. *J Usability Stud* 2008;**4**:1–6.

Address for correspondence

Danny T.Y. Wu, PhD, MSI, FAMIA
Assistant Professor, Biomedical Informatics & Pediatrics
Univerisity of Cincinnati College of Medicine
231 Albert Sabin Way, ML0840
Cincinnati OH, 45229, USA
Office: +1 (513) 558-6464
Email: wutz@ucmail.uc.edu