The Typology of Digital Health Apps According to their Quality Scores and User Ratings: K-Means Clustering

Maciej Hyzy
Ulster University
maciejmarekzych@gmail.com

Raymond Bond
Ulster University
rb.bond@ulster.ac.uk

Maurice Mulvenna
Ulster University
md.mulvenna@ulster.ac.uk

Lu Bai
Ulster University
l.bai@ulster.ac.uk

Simon Leigh
University of Warwick, ORCHA
simon.leigh@orcha.co.uk

ABSTRACT

This study focuses on discovering the types of digital health apps that exist in accordance with several quality characteristics and their user ratings on app stores. The quality scores include scores for the app's user experience (UX), its data privacy (DP) and professional clinical assurance (PCA) which are scores provided by ORCHA that use many objective questions to quality assess health apps (ORCHA stands for The Organisation for the Review of Care and Health Apps). K-means clustering has been used to group many digital health apps (n>1700) that have similar traits. We describe 6 different types of digital health apps. This study shows that one cluster (or type) comprise of 23.8% of health apps which typically have good user ratings and high-quality scores. Another cluster of apps comprise of 27.2% of health apps, which typically have low PCA scores but high UX and DP scores with good user ratings, indicating that this cluster of health apps are held back by their PCA score from becoming 'the highest quality' health apps.

CCS CONCEPTS

• **Applied computing**; • **Health informatics**;

KEYWORDS

mHealth quality traits, cluster analysis, digital health apps

ACM Reference Format:

Maciej Hyzy, Raymond Bond, Maurice Mulvenna, Lu Bai, and Simon Leigh. 2023. The Typology of Digital Health Apps According to their Quality Scores and User Ratings: K-Means Clustering. In *European Conference in Cognitive Ergonomics (ECCE '23), September 19–22, 2023, Swansea, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3605655.3605786

1 INTRODUCTION

This study has been conducted in cooperation with the Organisation for the Review of Care and Health Apps (ORCHA), a UK based digital health compliance company. ORCHA has used their tool, ORCHA Baseline Review (OBR) [6], to assess the quality (quality defined as "compliance with best practice standards") of over 1700 digital health apps that have been used in this study. Many health apps in use today may not be of sufficient quality. For example, current research suggests that health apps for the treatment of mental health conditions may have poor data governance and data sharing practices, and possibly harmful content [1] [3]. To mitigate risks associated with digital health apps, they need to be quality assured [4]. The objective of this study is to uncover similarities and differences in traits among digital health apps regarding characteristics related to the quality of the digital health apps and their user ratings on the app stores. Uncovering similarities and differences of traits via k-means cluster analysis can indicate areas where digital health apps can improve regarding quality assessment and inform on the state of digital health apps today. The size of cluster will indicate the prevalence of these traits amongst health apps. The use of k-means clustering will also provide a typology to help describe the types of health apps that exist in accordance with characteristics related to quality factors and user rating.

2 METHODS

2.1 The secondary dataset

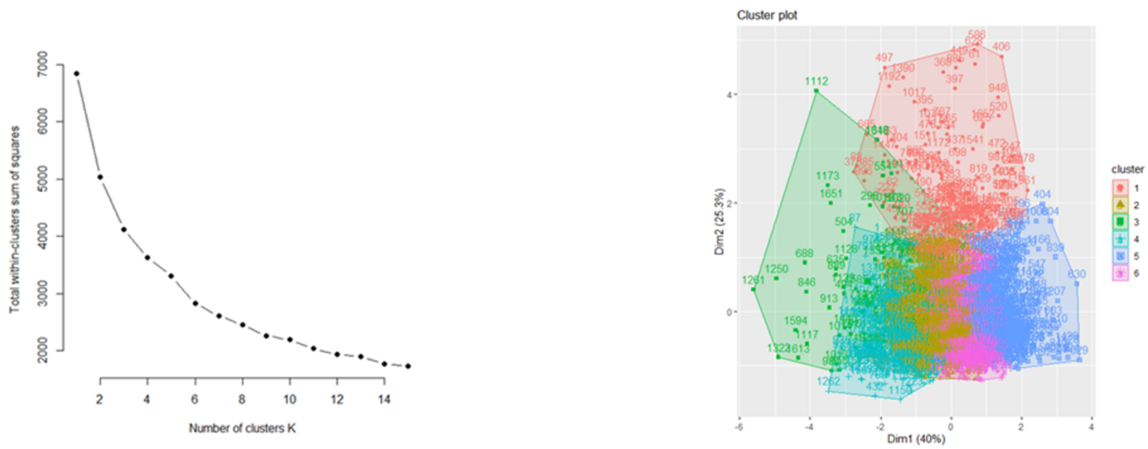
ORCHA dataset consists of 1712 digital health apps that have been quality assessed with OBR and rated by users. For 310 apps both Android and iOS version have been counted as separate apps, resulting in 620 assessments. OBR consists of three sections professional/clinical assurance (PCA), user experience (UX) and data privacy (DP). Each app has been assessed by two ORCHA reviewers where in the case of a dispute a third reviewer would be involved to resolve dispute.

2.2 Statistical analysis

R studio and R programming language has been used to conduct the analysis and produce figures. Elbow method has been used to determine the optimal number of clusters for the analysis. Mean and standard deviation (SD) have been calculated for user rating and the scores for reference. Shapiro-Wilk test has been used to check if the user ratings or the scores are normally distributed. Following results of the Shapiro-Wilk test, the unpaired two-samples Wilcoxon test has been used to compare corresponding user ratings and the scores among clusters, to check for statistical significance. P-value of .05 has been considered statistically significant.

ECCE '23, September 19–22, 2023, Swansea, United Kingdom

ACM ISBN 979-8-4007-0875-6/23/09.
https://doi.org/10.1145/3605655.3605786



a) b)

Figure 1: a) Elbow method for selecting number of clusters b) K-means with 6 clusters on user rating, PCA, UX and DP scores.

Table 1: Cluster numbering, labelling and description.

Cluster number	Cluster label	Description
1	Lower user rating	These are the apps that have low user rating but intermediate on the UX, PCA and DP scores
2	Lower PCA	These are the apps with low PCA score but high on the UX and DP scores
3	Lower scores	These are the apps with low UX, PCA and DP scores but high user rating
4	Lower PCA/DP	These are the apps with low PCA and DP scores, but high UX score and user rating
5	All high	These are the apps with high UX, PCA and DP scores and high user rating
6	Lower DP	These are the apps with intermediate DP score, high UX and PCA scores and high user rating

2.3 Consent

This secondary data analysis study gained ethical approval by Ulster University (ethics filter committee, Faculty of Computing, Engineering and the Built Environment). The developers under consideration provided implicit consent for use of their data for research purposes. All reviews, unless explicitly asked to be removed by the developer, are covered as suitable for research in ORCHA's privacy policy [5].

3 RESULTS

Figure 1 depicts the elbow method used to determine number of cluster and visual representation of 6 clusters. 6 cluster have been chosen due to the amount of variability that it explains.

After conducting cluster analysis and examining the centers, the following labels have been assigned to the clusters: Cluster 1 – Lower user rating, Cluster 2 – Lower PCA, Cluster 3 – Lower scores, Cluster 4 – Lower PCA/DP, Cluster 5 – All high and Cluster 6 – Lower DP. Details are shown in table 1.

Table 2 depicts cluster centers for each cluster across user rating, PCA, UX and DP. The scores range was from 0 to a 100, user rating range was from 1 to 5. The clusters show that different health apps pose different traits. Mean and standard deviation (SD) of the data has been depicted in the table for reference.

Shapiro-Wilk test indicated that user ratings and all scores are not normally distributed (P<.001, for all). Hence, unpaired two-samples Wilcoxon test has been used to check if center values are statistically significantly different for user-ratings and each of the scores among clusters as shown in tables 3, 4, 5, 6. **Non-statistically** significant results have green background. Traffic light color coding was used where background with high P-values were colored green.

Figure 2 depicts boxplots for user ratings and the scores against each of the 6 clusters

4 DISCUSSION

There are more than 350,000 digital health apps on the market today [2], understanding their traits with cluster analysis can be a useful way of identifying areas where these apps could be improved regarding their quality. Figure 1a indicates that 6 clusters is a good choice to conduct k-means cluster analysis. Figure 1b shows how different apps (represented by numbers), have been assigned into different clusters. The center values of these clusters can be seen in table 2. The results indicate that around 23.8% of health apps have good user ratings and high scores. Cluster 'Lower PCA' indicates that for 27.2% of the health apps PCA scores are low but UX and DP scores are high with good user rating. Indicating that health

Table 2: Cluster analysis on 1712 health apps. Traffic light color coding where scores <51 have red background, orange between 51 and <65, green for 65+. For user rating <2 is red, between 2 and <4 is orange, and 4+ is green.

		Cluster centers					
Variables	**Mean (SD)**	Lower user rating	Lower PCA	Lower scores	Lower PCA/DP	All high	Lower DP
User rating	4.26 (.694)	2.767080	4.443065	4.091852	4.438005	4.430508	4.548045
PCA score	53.2 (24.8)	56.37953	32.20638	40.43941	31.89225	77.25283	70.00369
UX score	74.8 (7.92)	75.08846	73.03326	50.57104	72.99388	79.76316	77.49129
DP score	63.7 (14.4)	64.37573	68.77163	56.24097	41.19157	76.73447	58.82744
	Mean ORCHA score (SD)	63.7 (12.4)	54.2 (5.93)	47.8 (12.8)	46.1 (7.13)	77.8 (6.22)	68.9 (5.50)
	Cluster size	184 (10.7%)	466 (27.2%)	73 (4.26%)	248 (14.5%)	408 (23.8%)	333 (19.5%)

Table 3: Unpaired two-samples Wilcoxon test for user rating scores, p-values.

	Lower user rating	**Lower PCA**	**Lower scores**	**Lower PCA/DP**	**All high**
Lower user rating					
Lower PCA	<.001				
Lower scores	<.001	<.001			
Lower PCA/DP	<.001	0.9277	<.001		
All high	<.001	0.7998	<.001	0.7858	
Lower DP	<.001	<.001	<.001	0.002045	<.001

Table 4: Unpaired two-samples Wilcoxon test for PCA scores, p-values.

	Lower user rating	**Lower PCA**	**Lower scores**	**Lower PCA/DP**	**All high**
Lower user rating					
Lower PCA	<.001				
Lower scores	<.001	0.07661			
Lower PCA/DP	<.001	0.1246	0.02746		
All high	<.001	<.001	<.001	<.001	
Lower DP	<.001	<.001	<.001	<.001	<.001

Table 5: Unpaired two-samples Wilcoxon test for UX scores, p-values.

	Lower user rating	**Lower PCA**	**Lower scores**	**Lower PCA/DP**	**All high**
Lower user rating					
Lower PCA	<.001				
Lower scores	<.001	<.001			
Lower PCA/DP	<.001	0.8836	<.001		
All high	<.001	<.001	<.001	<.001	
Lower DP	<.001	<.001	<.001	<.001	<.001

apps are held back by their PCA score from becoming 'The highest quality' health apps, the similarities among clusters can also be seen in figure 2. The results of this analysis indicate that user ratings are not an indication of quality scores (PCA, UX and DP), as indicated by clusters 'Lower rating', 'Lower scores' and 'Lower PCA/DP'. Health apps can receive decent scores but be rated poorly by users (cluster 'Lower rating') or can receive high user rating but score poorly (clusters 'Lower scores' and 'Lower PCA/DP'). The unpaired two-samples Wilcoxon test in tables 3, 4, 5, 6 shows that some of the scores are not statistically significantly different. For UX cluster 'Lower PCA' with 'Lower PCA/DP', for PCA cluster 'Lower PCA' with 'Lower scores', and 'Lower PCA/DP', for DP cluster 'Lower scores' with 'Lower DP', and for user rating cluster 'Lower PCA'

Table 6: Unpaired two-samples Wilcoxon test for DP score, p-values.

	Lower user rating	Lower PCA	Lower scores	Lower PCA/DP	All high
Lower user rating					
Lower PCA	<.001				
Lower scores	<.001	<.001			
Lower PCA/DP	<.001	<.001	<.001		
All high	<.001	<.001	<.001	<.001	
Lower DP	<.001	<.001	0.4771	<.001	<.001

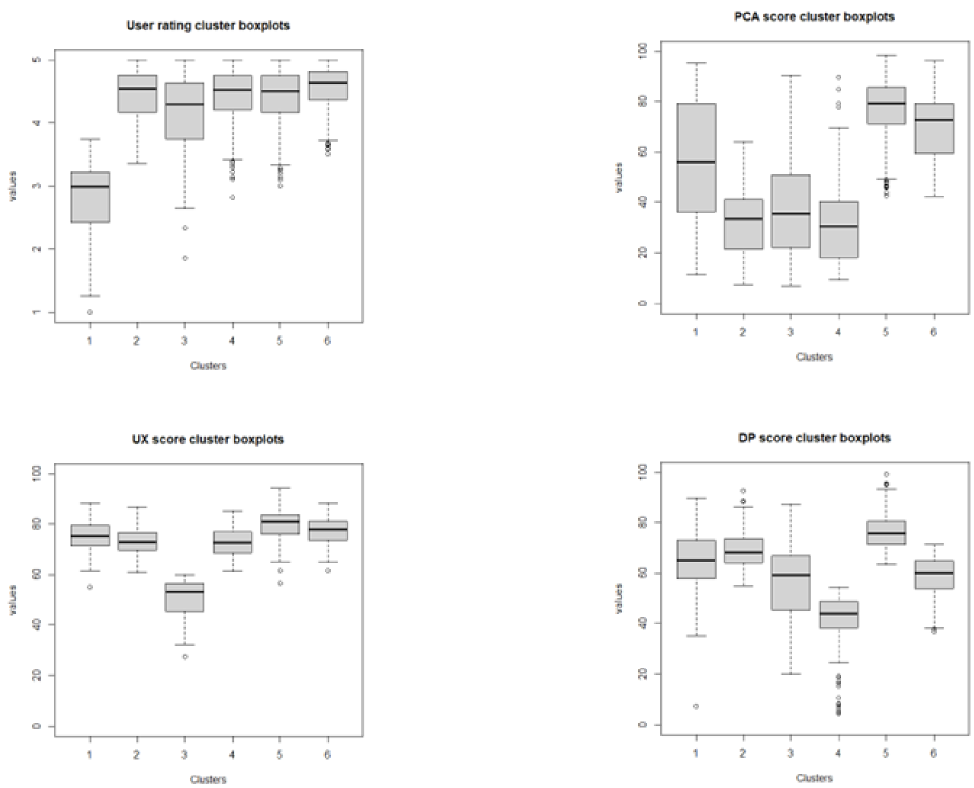


Figure 2: Boxplots for usr rating, PCA, UX and DP scores per each cluster. The following labels have been assigned to the clusters: Cluster 1 – Lower user rating, Cluster 2 – Lower PCA, Cluster 3 – Lower scores, Cluster 4 – Lower PCA/DP, Cluster 5 – All high and Cluster 6 – Lower DP.

with 'Lower PCA/DP' and 'All high', and 'Lower PCA/DP' with 'All high'.

ACKNOWLEDGMENTS

We'd like to acknowledge the contribution of the reviewers who have used the ORCHA assessment tool to assess digital health apps, making this study possible. We'd also like to thank digital health app developers who have allowed their app to be assessed for the purposes of research.

Funding: This study has been funded by CAST award / DfE and ORCHA.

REFERENCES

[1] Huckvale, K. *et al.* 2020. Smartphone apps for the treatment of mental health conditions: status and considerations. Current Opinion in Psychology. 36, (Dec. 2020), 65–70. DOI:https://doi.org/10.1016/J.COPSYC.2020.04.008.
[2] Kern, J. *et al.* 2021. written consent of IQVIA and the IQVIA Institute. Digital Health Trends. (2021).
[3] Larsen, M.E. *et al.* 2016. A Systematic Assessment of Smartphone Tools for Suicide Prevention. PLoS ONE. 11, 4 (Apr. 2016). DOI:https://doi.org/10.1371/JOURNAL.PONE.0152285.
[4] Mathews, S.C. *et al.* 2019. Digital health: a path to validation. npj Digital Medicine 2019 2:1. 2, 1 (May 2019), 1–9. DOI:https://doi.org/10.1038/s41746-019-0111-3.
[5] ORCHA Privacy Policy: https://appfinder.orcha.co.uk/privacy-policy/. Accessed: 2022-08-11.
[6] Review Documentation - Review Development & Resources | Exte: https://confluence.external-share.com/content/b6055aac-83e4-4947-be0e-ebb8c39559ef. Accessed: 2022-03-13.