Qualitative Assessment of Health Apps: An Emerging Market Study

Garima Goyal\textsuperscript{1*} and Sonika Sharma\textsuperscript{1}

\textsuperscript{1}Department of Food and Nutrition, Punjab Agricultural University, Ludhiana, India.

\textbf{Authors' contributions}

This work was carried out in collaboration between both authors. Author GG managed the literature searches, collected data and wrote the first draft of the manuscript. Author SS was responsible for conceiving the idea, designed the study and edited the manuscript. Both the authors read and approved the final manuscript.

\textbf{Article Information}

DOI: 10.9734/CJAST/2020/v39i4831237

Editor(s):
(1) Dr. Ming-Chih Shih, Chinese Culture University, Taiwan.

Reviewers:
(1) Federico Diano, University of Naples “Federico II”, Italy.
(2) Iffah Syafiqah binti Meor, International Islamic University Malaysia, Malaysia.

Complete Peer review History: http://www.sdiarticle4.com/review-history/64936

\textbf{ABSTRACT}

\textbf{Aims}: Increased usage of health apps has led to need for their quality assessment for safeguarding interest of various stakeholders. This study attempts to undertake qualitative assessment of health apps in an emerging market, India.

\textbf{Study Design}: Health apps were evaluated by the experts and secondary data was used for rating of the health apps. Indian food data base was used for evaluating content accuracy.

\textbf{Place and Duration of Study}: Department of Food and Nutrition, College of Community Science, Punjab Agricultural University, Ludhiana, between January 2019 and August 2019.

\textbf{Methodology}: Top 10 health apps, identified from response of 400 users, are assessed qualitatively by expert raters using App Quality Evaluation Questionnaire. Content accuracy in terms of macronutrient measurements is assessed using Mean Absolute Percent Error (MAPE). Relationship between average user ratings and various aspects of qualitative assessment is explored using linear regression.

\textbf{Results}: Majority of the apps performed well in terms of functionality, interactivity, security and aesthetics. Relatively poor performance is observed in terms of accountability, behavior change techniques and scientific coverage. Regression analysis indicate that Functionality ($p=0.035$) and engagement ($p=0.024$) features significantly influence user ratings and overshadow scientific coverage and accuracy ($p=0.798$). MAPE values indicate considerable variations from Indian food data base across the apps especially in terms of protein and energy.

*Corresponding author: E-mail: garima.goyal.goyal.15@gmail.com;
Conclusion: Quality assessment of top 10 health apps, in Indian scenario, indicates that the apps are proficient on functionality. At the same time, the apps fair poorly in terms of scientific content and accountability. There is a pertinent need for statutory regulations as well as voluntary efforts for improving the scientific content. Collaborative efforts of app developers with scientific institutions should be promoted.

Keywords: mhealth, smartphone, health apps, content analysis, quality assessment.

1. INTRODUCTION

Recent years have witnessed a sharp increase in the usage of mobile phones especially among adults, adolescents and children [1-3]. Young adults are the most skilled amongst all age groups in using mobile phones. In India, the prevalence of mobile phone usage is the highest with 79.2 percent in 18-30 years age group in India [4]. Mobile data traffic is expected to grow a seven-fold from 2016 to 2021, at a compound annual growth rate of 47% globally and 49% in India [5]. This technology continues to transform the way users accomplish various daily life activities, including those associated to sustaining health and wellness. Health-based apps are recognized as the third fastest growing app category online, subsequent to games and utilities [6]. More than 100,000 health apps are accessible in the health and fitness category of iTunes and Google Play store [4]. Smartphone health apps are promising in terms of augmenting weight management as well as behavioral changes amongst users [7-12]. It is essential to assess the quality of these apps for ensuring progressive health outcomes. World Health Organization (WHO) reported an increase in number of mHealth initiatives by the developed countries relative to the developing nations [13-15].

1.1 Objective

Worldwide, evaluation of health apps remains a lesser treaded domain and more so in context of developing economies. In general, there is a scarcity of research studies assessing the quality and effectiveness of health apps and consequent regulatory implications. Studies involving quality evaluation of health apps in Indian context are rare to find. Present study primarily deals with the qualitative assessment of top health apps being used in Indian scenario. It also evaluates content accuracy of the apps operational in India on basis of comparing macronutrient measurements with the local Food Database. Further, the study also explores the relationship between qualitative assessment given by the experts with the user rating for the health apps.

2. METHODOLOGY

For the purpose of qualitative assessment and content analysis, we identified top 10 health apps by surveying 400 university students of Punjab Agricultural University, Ludhiana, in the business capital of Indian state of Punjab. We used a scoring scale, i.e. App Quality Evaluation Questionnaire (AQEQ) for the assessment of app quality based on previous researches [16-19]. The scale consist of 124 points, out of which 95 points are for quality assessment features like phenotype assessment, accountability, Behaviour Change Techniques (BCTs), scientific coverage and accuracy, accessibility, inclusion of technology enhanced features, security, interactivity and connectivity features of the app. Scoring involved awarding a score of ‘1’ for an app with specified characteristics and ‘0’ otherwise. The remaining 29 measures were evaluated on the basis of Mobile App Rating Scale (MARS) [18]. A 5-point rating scale was used for this purpose. Four independent raters assessed the identified to ten apps qualitatively on basis of various parameters included in AQEQ. These raters were skilled in using health apps and were well acquainted with domain of food and nutrition. For the purpose of evaluating content accuracy, we compared the Food Data Base (FDB) of each app with an Indian Standard ‘Indian Food Composition Tables 2017’, proposed by National Institute of Nutrition, Indian Council of Medical Research (ICMR) [20]. We checked and compared the BMI calculated by the app with WHO cut-off for overweight [21]. We attempt to ascertain the qualitative measures of health apps having significant influence on user rating. Using linear regression, we explored the relationship between various qualitative assessment measures rated by the raters and app ratings determined by the users.

2.1 Statistical Analysis

We evaluated content accuracy using Mean Absolute Percent Error (MAPE). MAPE carries the property of scale neutrality and is easily interpretable. Higher value of MAPE is indicative of poorer content accuracy.
\[ MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{R_i - C_i}{R_i} \right| \]

where \( R_i \) is the reference value of macronutrient from Indian Food Database and \( C_i \) is the calculated value from app database.

We take average user rating (available at Playstore) as dependent variable and run regression analysis by putting various aspects of qualitative assessment such as accountability, Behaviour Change Techniques (BCTs), scientific coverage and accuracy, accessibility, inclusion of technology enhanced features, security, interactivity etc as independent variables.

We used Cohen’s kappa (\( \kappa \)) [22] to measure interrater reliability [23-24].

\[ \kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \]

where:

\( p_o \) = the relative observed agreement among raters.

\( p_e \) = the hypothetical probability of chance agreement.

For various pairs of raters, calculated values for Cohen’s Kappa had a range from 0.730 to 0.805. Kappa coefficient between 0.61-0.80 indicates a substantial agreement [22]. Minimum value of Cohen’s Kappa (\( K=0.730, \ P<0.0001 \)) is statistically significant. We use SAS 9.4 for statistical analysis.

3. RESULTS

Findings based on various dimensions recorded using AQEQ are presented in this section.

3.1 Qualitative Assessment of Health Apps

Table 1 summarizes component wise evaluation for various apps. Overall, HealthifyMe and My Fitness Pal are the top apps with score of 94.60 and 82.65 out of maximum of 124.

3.1.1 Phenotype

The phenotype feature comprise of age and gender of the user. All apps considered phenotype dimensions except for iPhone Health app.

3.1.2 Accountability

Accountability measures evaluate an app’s authorship (credentials and affiliations), attribution (provision of information sources and references) and sponsorship disclosure. Out of 10 apps, 7 apps scored zero on the basis of accountability. Samsung Health has the highest accountability score of 4 (out of maximum of 5). HealthifyMe disclosed sponsorship and Google Fit detailed references and information sources.

3.1.3 Scientific coverage and content

Scientific coverage and content accuracy investigates the range and accuracy of information related to weight management and general dietary advice provided by the apps. We consider various characteristics such as nutritional assessment, anthropometric assessment, Body Mass Index (BMI), calorie counter and physical activity recommendations. Overall the selected health apps scored 17.35 out of 40 for scientific coverage and accuracy. HealthifyMe received the highest score of 35. My Fitness Pal and Samsung Health app have scores of 26 and 22.5 respectively. All the apps considered the measures of current height and weight; however, target weight was set up by 7 apps only. Only 2 apps, took waist circumference and hip circumference into account. BMI was calculated by only 4 apps out of which only two apps were recommending realistic weight loss goal i.e. 0.5-1 kg weight loss/ week. Eighty percent of the apps offered features of logging Physical Activity (PA), tracking PA, and recommended PA. Only two apps considered recommending PA on the basis of energy consumed and spent.

3.1.4 Technology-enhanced features

The apps with the relatively higher inclusion of technology-enhanced features are My Fitness Pal and Fitbit, with maximum score of 12. The most frequent theoretical background or strategies of the apps were monitoring/ tracking, assessment, feedback, and information/ education.

3.1.5 Incorporation of the behaviour change techniques (BCTs)

There is a substantial variation in the numbers of BCTs present in various apps, with an overall average of 12.42 techniques per app (out of maximum of 26). HealthifyMe had the highest
incorporation of the BCTs (19.75), followed by Google Fit and Nike Run Club (15.75).

3.1.6 Interactivity

The content analysis of apps revealed that a broad array of features are available within apps, including pairing with other devices via Bluetooth to measure physical activity.

3.1.7 Engagement

The engagement feature marked the app for fun, interesting, customizable, interactive features (send alerts and reminders) and see how well the app is targeted to audiences. HealthifyMe had the highest engagement score of 4.30 (out of maximum of 5).

3.1.8 Functionality

It accounts for app functioning in terms of ease to use, navigation, flow logic, and gestural design of the app. On the basis of functionality, the apps had mean score of 3.06 (out of maximum of 4). My Fitness Pal app got the highest score of 3.70, closely followed by Fitbit (3.65) and HealthifyMe (3.55).

3.1.9 Aesthetics

Features like graphic design, overall visual appeal, color scheme and stylistic consistency are considered for aesthetics feature assessment. HealthifyMe attained the highest score of 2.85, followed by iPhone Health app (2.8) out of a maximum of 5.

3.1.10 Information

This aspect deals with quality information (e.g. text, feedback, measures, references) from a credible source. HealthifyMe scored the highest score of 4.90 (out of maximum of 7).

3.1.11 App subjective quality

The apps had a quality score ranging from 0.95 to 2.45 out of a maximum of 4, with overall mean score of 1.85. Home Workout- No Equipment had the highest score of 2.45, followed by HealthifyMe (2.30) and Samsung Health app (2.25).

3.1.12 App-specificity

This feature is used to perceive the impact of the app on the user’s knowledge, attitudes, intentions to change as well as the likelihood of actual change in the target behavior. Overall mean score of app specificity of all the apps was 2.93.

Overall, the selected health apps performed better in terms of security (95 percent), interactivity (90 percent), functionality (76.5 percent) and aesthetics (75.33 percent). On the other hand, these apps are lacking on basis of accountability (14 percent), scientific coverage and content accuracy along (43.38 percent) with Behavior Change Techniques (47.80 percent).

3.2 Content Accuracy Based on Macronutrient Measurements

We evaluate content accuracy of Food Database (FDB) in these health apps by comparing the macronutrient measurements provided by the apps with that of Indian Food Composition Tables. Table 2 presents MAPE values. On the basis of MAPE, the highest errors are found in energy content (53.16±16.70 for Papaya) followed by proteins (42.06±1.37 for Papaya). Relatively larger MAPE values for protein are found as compared to other macronutrients.

3.3 Relationship between User Ratings and Qualitative Assessment Measures

We used regression analysis to evaluate the relationship between user ratings in the commercial app store and quality assessment measures used in the study. We take user ratings as dependent variable and various aspects of qualitative assessment made by the raters as independent variables. Table 3 presents the findings of various regression equations. Table 3 reveals a significant positive relation between user rating and incorporation of BCTs (P=.02), engagement features (P=.03), functionality features (P=.04), quality of information provided by the app (P=.04), app subjective quality (P=.02) and app specificity (P=.03). On the other hand, there is no significant association of user rating with app quality on features like accountability, scientific coverage and content accuracy, accessibility, technology enhanced features, interactivity, aesthetics and total quality assessment score.

4. DISCUSSION

In context of the present study, Qualitative Assessment of Health apps reveal a variety of
issues. Firstly, the findings of the study indicate that the majority of the apps are effective in terms of functionality. But at the same time, the apps also suffer from the issues related to poor accountability and scientific coverage which parallels the findings in Australia and Korea [25-26]. Generally speaking, the finding related to poor accountability assumes greater significance in light of limited ability of the users to assess these apps on the basis of scientific coverage. Trust is considered as an important dimension for adoption and sustenance of new technology. Given the poor quality in terms of accountability and scientific coverage, there could be a possible case of trust erosion. Secondly, the findings related to relationship between user rating and app quality dimensions, based on regression analysis, throw up vital insights regarding the user assessment of these apps. Findings indicate that crucial dimensions of app quality such as accountability, scientific coverage and content accuracy are not having significant relationship with the user rating. Significant determinants of user rating are functionality, subjective quality and information. Technically speaking, accountability and scientific coverage should also be the important determinants of app quality. This phenomenon may be on account of poor awareness and knowledge on part of users as well as inability of the app providers to highlight these important dimension. Given the case, there is a high probability of subjective factors and marketing efforts winning over the content accuracy and accountability. There is a clear case of the majority of the apps lacking scientific support. A possible explanation for the lack of scientific support could be that most of the apps are commercial. Therefore, there is a need to promote the development of apps in collaboration with scientific institutions [27]. Previous researches have highlighted the lack of theoretical content present in interactive technologies in apps [28-30] intended to promote health behaviour change [12,23,31-33]. Consistent with previous research [29], our findings demonstrate the relative absence of behaviour change strategies present in physical activity and dietary apps. Thirdly, in terms of content accuracy based on measurement of macronutrients, considerable variations from FDB are found for the majority of the apps. These variations are relatively prominent in case of Protein and Energy calculations. There is an obvious need on part of the app providers to improve the content accuracy. Banerjeeet al has reported a deficiency of common Indian recipes in FDB of these apps [34]. This phenomenon may be on account of app development taking place in developed parts of the world and the variations of cuisines across the globe. Given the findings, there is a definite case for enhanced efforts on the part of the app developers to go local in terms of including local recipes and using local food database. Overall, accountability and scientific coverage remain the major areas of concern regarding these apps. Given the limitations on the part of users to assess these apps, there is a pertinent need for regulatory interventions for the health apps. Previously, lay users and dietitians have communicated similar opinions over the integrity, inclusiveness, accuracy, and general quality of health and fitness apps, as well as the apps.

<table>
<thead>
<tr>
<th>Features</th>
<th>Max score</th>
<th>Overall Mean± SD</th>
<th>Performance (% age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>2</td>
<td>1.90±0.32</td>
<td>95.00</td>
</tr>
<tr>
<td>Accountability</td>
<td>5</td>
<td>0.70±1.34</td>
<td>14.00</td>
</tr>
<tr>
<td>Behavior change techniques</td>
<td>26</td>
<td>12.43±5.82</td>
<td>47.80</td>
</tr>
<tr>
<td>Scientific coverage and content accuracy</td>
<td>40</td>
<td>17.35±9.04</td>
<td>43.38</td>
</tr>
<tr>
<td>Accessibility</td>
<td>5</td>
<td>3.05±1.06</td>
<td>61.00</td>
</tr>
<tr>
<td>Technology enhanced features</td>
<td>12</td>
<td>6.70±3.58</td>
<td>55.83</td>
</tr>
<tr>
<td>Security</td>
<td>3</td>
<td>3.00±0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Interactivity</td>
<td>2</td>
<td>1.80±0.63</td>
<td>90.00</td>
</tr>
<tr>
<td>Engagement</td>
<td>5</td>
<td>3.24±6.00</td>
<td>64.80</td>
</tr>
<tr>
<td>Functionality</td>
<td>4</td>
<td>3.06±0.52</td>
<td>76.50</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>3</td>
<td>2.26±0.18</td>
<td>75.33</td>
</tr>
<tr>
<td>Information</td>
<td>7</td>
<td>3.63±0.68</td>
<td>51.86</td>
</tr>
<tr>
<td>App subjective quality</td>
<td>4</td>
<td>1.85±0.52</td>
<td>46.25</td>
</tr>
<tr>
<td>App specificity</td>
<td>6</td>
<td>2.93±0.77</td>
<td>48.83</td>
</tr>
<tr>
<td>Total</td>
<td>124</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Overall mean absolute percent error (MAPE) for macronutrient measurements

<table>
<thead>
<tr>
<th>Food item</th>
<th>Protein Mean± SD</th>
<th>Fat Mean± SD</th>
<th>Carbohydrate Mean± SD</th>
<th>Energy Mean± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wheat Flour</td>
<td>10.69±6.55</td>
<td>0.16±0.09</td>
<td>0.12±0.09</td>
<td>9.18±4.69</td>
</tr>
<tr>
<td>Red Kidney Beans</td>
<td>34.57±36.24</td>
<td>1.19±1.47</td>
<td>0.43±0.37</td>
<td>31.14±30.17</td>
</tr>
<tr>
<td>Soybean</td>
<td>33.22±22.39</td>
<td>10.69±16.83</td>
<td>1.45±0.28</td>
<td>6.86±4.68</td>
</tr>
<tr>
<td>Spinach</td>
<td>29.60±26.11</td>
<td>0.33±0.21</td>
<td>0.50±0.11</td>
<td>27.44±37.99</td>
</tr>
<tr>
<td>Tomato</td>
<td>36.67±55.08</td>
<td>0.45±0.29</td>
<td>0.41±0.03</td>
<td>6.31±3.70</td>
</tr>
<tr>
<td>Apple</td>
<td>19.54±14.36</td>
<td>0.43±0.31</td>
<td>0.04±0.01</td>
<td>12.83±6.42</td>
</tr>
<tr>
<td>Papaya</td>
<td>42.06±1.37</td>
<td>0.19±0.18</td>
<td>0.94±0.00</td>
<td>53.16±16.70</td>
</tr>
<tr>
<td>Pineapple</td>
<td>38.72±47.73</td>
<td>0.57±0.37</td>
<td>0.29±0.03</td>
<td>11.36±4.66</td>
</tr>
<tr>
<td>Tomato</td>
<td>12.99±14.64</td>
<td>3.34±5.26</td>
<td>0.33±0.08</td>
<td>31.44±22.60</td>
</tr>
<tr>
<td>Walnut</td>
<td>3.62±2.73</td>
<td>15.64±27.08</td>
<td>0.26±0.04</td>
<td>2.20±0.46</td>
</tr>
<tr>
<td>Milk (Cow)</td>
<td>4.70±3.38</td>
<td>2.30±3.75</td>
<td>0.06±0.00</td>
<td>13.58±7.64</td>
</tr>
<tr>
<td>Egg Whole</td>
<td>4.22±3.54</td>
<td>6.43±10.68</td>
<td>0.00±0.00</td>
<td>17.71±9.57</td>
</tr>
<tr>
<td>Mustard Oil</td>
<td>0.00±0.00</td>
<td>0.00±0.00</td>
<td>0.00±0.00</td>
<td>1.73±0.09</td>
</tr>
</tbody>
</table>

Table 3. Relationship between the user ratings and quality assessment measures

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Intercept</th>
<th>Coefficient (B)</th>
<th>P-value</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>4.309</td>
<td>.088</td>
<td>.49</td>
<td>0.059</td>
</tr>
<tr>
<td>Behavior change techniques</td>
<td>3.612</td>
<td>.061*</td>
<td>.02</td>
<td>0.544</td>
</tr>
<tr>
<td>Scientific coverage and content accuracy</td>
<td>4.276</td>
<td>.005</td>
<td>.78</td>
<td>0.010</td>
</tr>
<tr>
<td>Accessibility</td>
<td>3.939</td>
<td>.146</td>
<td>.44</td>
<td>0.078</td>
</tr>
<tr>
<td>Technology enhanced features</td>
<td>4.332</td>
<td>.006</td>
<td>.91</td>
<td>0.002</td>
</tr>
<tr>
<td>Interactivity</td>
<td>4.900</td>
<td>-.294</td>
<td>.27</td>
<td>0.150</td>
</tr>
<tr>
<td>Engagement</td>
<td>2.655</td>
<td>.516*</td>
<td>.03</td>
<td>0.490</td>
</tr>
<tr>
<td>Functionality</td>
<td>2.277</td>
<td>.645*</td>
<td>.04</td>
<td>0.444</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>2.042</td>
<td>.915</td>
<td>.30</td>
<td>0.136</td>
</tr>
<tr>
<td>Information</td>
<td>2.869</td>
<td>.404*</td>
<td>.04</td>
<td>0.442</td>
</tr>
<tr>
<td>App subjective quality</td>
<td>3.079</td>
<td>.699*</td>
<td>.02</td>
<td>0.504</td>
</tr>
<tr>
<td>App specificity</td>
<td>3.081</td>
<td>.426*</td>
<td>.03</td>
<td>0.460</td>
</tr>
<tr>
<td>Total quality assessment score</td>
<td>3.639</td>
<td>.011</td>
<td>.19</td>
<td>0.209</td>
</tr>
</tbody>
</table>

reputation and legality of the app sources [35-36]. App developers act as the suppliers of information to the users and thus it is essential that they establish a standardized approach to deliver medical information and provide specific instructions for mobile devices [37]. Users invest monetary as well as non-monetary resources while using these apps. Therefore, to ensure fairness to the users, regulatory mechanism for the greater accuracy of these apps must be put in place. Given the rising health concerns and consciousness among the public, the usage of these apps is expected to increase in future. At the same time, users must be made aware and assured regarding the quality of these apps.

5. CONCLUSION

Increased presence and use of health apps necessitates their qualitative evaluation for the benefit of the stakeholders especially the users. Qualitative assessment of health apps indicates few areas of concerns such as accuracy, content quality and scientific coverage. Further, inaccuracies in the calculations of macronutrients are also observed. Functionality overshadows scientific accuracy for determining the user rating. There is a pertinent need for improving the scientific coverage and accuracy of the health apps both through voluntary initiatives as well as regulatory mechanisms. Future research may investigate the impact of health apps on health and physical well-being of the users through anthropometric and clinical studies.

CONSENT

As per international standard or university standard, Participants’ written consent has been collected and preserved by the authors.

COMPETING INTERESTS

Authors have declared that no competing interests exist.
REFERENCES


