Developing and testing a health app evaluation framework for organisations to recommend the best health apps for consumers in the Australian setting

EXECUTIVE SUMMARY

June 2022

Authors:

Sithara Wanni Arachchige Dona

Dieu Nguyen

Mary Rose Angeles

Paul Cooper

Natalie Winter

Mary Lou Chatterton

Anna Peeters

Martin Hensher

Deakin Health Economics, School of Health and Social Development
Institute for Health Transformation
Deakin University

INSTITUTE FOR HEALTH TRANSFORMATION

Acknowledgement of country

The researchers would like to acknowledge the Wadawurrung and Wurundjeri peoples of the Kulin Nation as the Traditional Owners of the lands on which we live and work. We pay our respect to their Elders past, present, emerging, and future.

Funding acknowledgement

This project was funded by Medibank Better Health Foundation (Medibank BHF). Professor Anna Peeters from Deakin University was supported by an NHMRC investigator grant.

Acknowledgement of involvement in the project

The researchers would like to acknowledge the Medibank staff's time and effort in working with the Deakin staff to test and improve the draft framework and advertise the survey on the Medibank website. In addition, the researchers are thankful to the focus group and survey participants for their time and valuable input.

Introduction

The number of health apps available on mobile devices continues to grow exponentially. However, there is very little authoritative guidance for consumers and health organisations to identify which of the health apps currently available in the market can be used safely and beneficially.

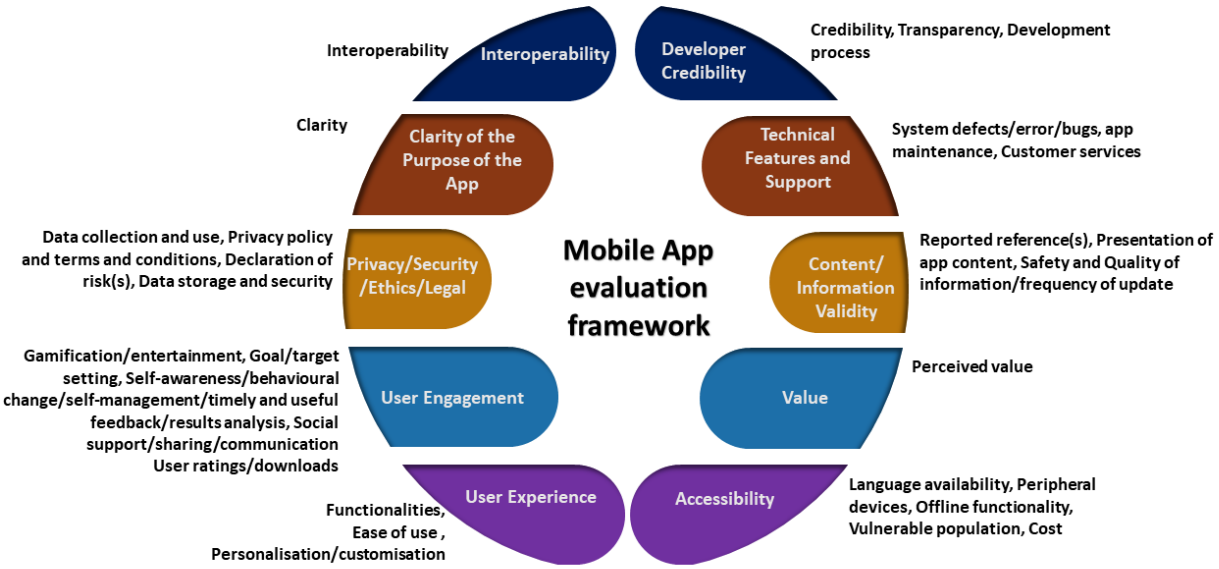
In this context, the aims of the project were:

1. Identify the priority domains from the scientific literature for evaluating digital health apps relevant to the Australian context.
2. Develop a framework based on the identified domains and domain items from the scientific literature.
3. Undertake health expert feedback on the identified domains and domain items' relevance, feasibility, and usability.
4. Evaluate the framework's usability when applied to apps covering general health issues and a range of common health issues or conditions frequently encountered in the general population.
5. Update the framework based on evaluation, expert and consumer feedback.

Phase I

Scoping Review: We conducted a scoping review of the literature to identify the priority domains most likely to be useful to include in a framework for evaluating and rating digital health apps. The review identified 97 studies that complied with a list of 430 assessment criteria. The most frequently used scaling mechanism was a 5-point Likert scale. Most studies have adopted summary statistics to generate the total scoring of each app, and the calculation of mean or average scores was the most popular approach. Some frameworks did not use any scaling or scoring mechanism and adopted criteria-based, pictorial, or descriptive approaches or "threshold" filter. Ten potential framework domains were identified across studies: *Clarity of Purpose of the App, Developer Credibility, Content/Information Validity, User Experience, User Engagement, Interoperability, Values, Technical Features and Support, Privacy/Security/Ethics/Legal, and Accessibility (Figure).* There are overlaps between some components of domains, suggesting there is some flexibility within frameworks and that categorisation of domains is not yet a standardised process. The selection of questions from the list of 430 assessment criteria, which was identified from our scoping review, for an app evaluation framework should be carefully conducted based on criteria including, but not limited to, the structure, the depth and the expected outcome from the question, and the subjectivity or objectivity, because individual perceptions of the quality of the questions can vary from one person to another. We published the scoping review in the Journal of the American Medical Informatics Association[1].

Domains identified from the scoping review



Focus Groups: We conducted four online focus group discussions with 17 health professionals and stakeholders with expertise in the digital health space regarding the relevance, feasibility, and usability of the identified domains from the scoping review. The majority of focus group participants (41%, n=7) were very familiar with a range of digital health apps on mobile devices, while only one participant (5%, n=1) had limited knowledge of a range of digital health apps on mobile devices Most of the focus group findings were similar to the scoping review findings. The other key findings, which we incorporated when drafting the framework, included the following.

The framework's perspective and the intended audience were deemed essential, especially when considering the domains' order or weight of importance. Regardless of who the intended audience is, *Privacy/Security/Ethical/Legal*, *Content/Information Validity*, and *Clarity of Purpose of the App* were felt to be the most important, while *Technical Features and Support* and *Interoperability* were seen as the least important. However, there was no definitive conclusion on the most critical domain.

For the *Accessibility* domain, health app evaluators should look beyond the language or online and offline functionality and review how the health apps extend or likely bridge the care for people with a disability or other vulnerable population groups.

On *Developer Credibility*, participants indicated that evaluators should consider the app maker's credibility and the factors driving the app content (i.e., source of funds or who commissioned the app). However, there were opposing views saying that the credibility of the maker should not be the basis of judging a quality app for the reason that credibility was easily manipulated or misleading.

User Engagement could be reviewed by checking the metrics or analytics of health apps, such as the number of downloads, prevalence of end-user visits, the frequency or constancy of usage, mobile traffic, and the attrition rate of end-users. Participants also recommended pilot testing and a literature search on measuring *User Engagement*, because there are no standardised approaches to measure this domain.

The *Value* domain should be clearly defined as it varies depending on the perspective of the individual user, organisation, or authorities. One participant said that assessing all the other domains would mean that *Value* of the app has been assessed. Several domains, such as the *Clarity of the Purpose of the App, User Engagement, User Experience,* and *Accessibility* domains were thought to be connected to the *Value* domain. Some participants noted that the app's price is not an indicator of the app's value to users.

Development of the Framework: The development of the framework was initiated in Phase I, and findings from the scoping review were used to develop the framework by synthesizing common domains and items. As the COVID-19 pandemic emerged, the preliminary framework (Version 1) was first used to assess the best COVID-19 apps in Australia. The framework Version 1 was tested on COVID-19 apps weekly from 24th March to 13th May 2020, which were available for download and usable in Australia. At the final test of COVID-19 apps, we examined 16 apps after several screening steps against the preliminary framework domains and scored them according to the rating matrix. Based on our evaluation as of 20 May 2020, we recommended '*Coronavirus Australia'*, '*Healthdirect'*, '*My Aus COVID 19*', and '*COVID Safe'* apps as being the highest-scoring COVID-19 apps then available in Australia. The *Value* domain was seen as hard to assess.

Phase II

Assessment of the framework: We further developed the framework by integrating the findings from the scoping review, the focus groups, and the preliminary testing on COVID-19 apps. The resulting Deakin Health E-technologies Assessment Lab Framework for Mobile Health Evaluation (D-HEAL FMHE) Version 2 consisted of 27 questions with 5-point numerical rating scale responses to each question/item grouped into ten domains. Version 1 was improved by expansion from seven to ten domains while updating the 5-point scale from a rating matrix of 'good, indicative good, average, indicative poor, and poor' to a numerical 5-point rating scale with answer options tailored for each question.

We then evaluated mental health apps using this revised D-HEAL FMHE. Multiple evaluators from Deakin University and Medibank assessed the framework's usability. Apps designed for use by people with depression and anxiety/stress apps (and which were available on both the Apple and Google Play stores) were identified as the priority area of interest for evaluation using the framework. Three raters evaluated 49 apps to select the top 10 based on average total scores. In terms of resourcing, the average time duration for screening, data collection, and evaluation per app was 3 minutes (171 seconds), 5 minutes (326 seconds), and 1 hour respectively.

Inter-rater reliability testing: The inter-rater reliability among the three raters' responses was statistically analysed using Fleiss's Kappa and Intraclass correlation coefficient (ICC). In Fleiss's Kappa, none of the apps was at the extremes of either poor agreement (below 0) or almost perfect agreement (0.81-1.0). 25 of the 49 apps (51%) showed moderate agreement (0.41-0.6), followed by fair agreement (0.21-0.40) for 17 apps (35%). The calculated Fleiss's Kappa for each question of the Framework (three raters and all the apps subject to each question) showed that the most frequent level of agreement among questions (n=13/27) was "slight" agreement among three raters. This indicated a need to review and revise the questions in the framework, which we acted upon at a later stage of the framework development.

For the ICC, the pooled analysis of all the apps (n=49) showed a high consistency of observation amongst raters and apps. ICC analysis for individual apps indicated that 61% of rated apps have high consistency (reliability) in rating (ICC, Alpha > 0.7, n= 30). Subgroup

analysis in each item and domain indicated the opportunity for improvement in questions 1-3, 9-13, 15, 17, 24, and 26. However, the issue of limited consistency could potentially be resolved by offering detailed training for questions 2-3, 13-15, and 17. Further refinement of questions 9-12 and 24-27 could also improve consistency. The domain *Privacy/Security/Ethics/Legal* consisted of questions 9-12 and had poorer average ICC in both poorly rated and highly rated apps. Due to the equivocal results for this domain, we will consider approaching this domain as an independent domain in our final version as a viable option to maintain the acceptable/good levels of consistency in other domains for the final scoring. At the same time, this domain remains a critical part of our framework.

Compared to Fleiss's Kappa analysis, the overall ICC result was consistent with some absolute agreement detected by app and item consistency rating tests. However, some apps were in disagreement and had a negative ICC score, suggesting sampling error attributed to a small sample size (i.e., low number of available raters).

Survey of the consumer experience of recommended apps: From the list of evaluated mental health apps, the top three scored apps ('ClearFear', 'Headspace', and 'Smiling Mind') were identified and communicated to the public by the Medibank website. Webpage visitors were invited to use these apps and participate in the survey study. We performed a pre-and post-survey of app users to collect consumers' experiences of using the top three mental health apps for depression and anxiety/stress, made available to the public by publishing them on the Medibank website along with a link to the survey. This survey aimed to investigate whether the consumer ratings on the apps correlated with the framework ratings on *User Engagement, User Experience* and *Value*. When they enrolled to participate in the study, participants were redirected to complete the baseline survey (S1). The first follow-up (S2) and the second follow-up (S3) were completed at the end of the first and fourth weeks, respectively.

The total number of survey participants was 55 at baseline, 27 at follow-up 1, and 23 at follow-up 3. Among those who had previously used mental health and wellbeing apps, the most frequently used app had been 'Smiling Mind' (n=15) followed by 'Headspace' (n=6) – both of which were also two of the three apps recommended by Medibank as part of this study.

Content/Information Validity was seen as the most important domain by all the participants and those who had previously used health apps, whereas *Interoperability* was ranked as the least important domain. This finding was consistent with the focus group findings. However, the rankings of those who had not previously used health apps differed significantly, with non-users ranking *Technical Features and Support* as the most important and *Developer Credibility* as the least important domains.

The mean ratings on the Likert scale for health app usability questions increased from S2 to S3. The change in the scale increased positively for 58% of the questions, which indicated that usability and satisfaction with the apps increased over time. All three apps had the same pattern of positive change. Respondents reported that their frequency of app use reduced overtime for all three apps. All the apps were rated as easy to use and easy to learn how to use, and navigation was convenient moving between screens. The Likert scale rating increased overtime for these features. App users were more satisfied with 'Smiling Mind' than 'Headspace', and our framework also rated these apps with a high score for user satisfaction related domains such as *User Engagement and User Experience*. 'Clear Fear' was the app ranked lowest for satisfaction by survey respondents, but the scoring from our framework indicated the opposite, where this app scored high for satisfaction related domains. It should be noted that only two app users used and completed the surveys on 'Clear Fear', which could be the reason for opposite results on the 'Clear Fear' app. Interestingly, 'Headspace' and 'Clear Fear' users reported a reduction in their rating of the app over time as being beneficial for their health and well-being, but the change was not statistically significant.

Revising the framework: We then incorporated the evaluator feedback, inter-rater reliability, and survey findings to revise and refine the framework (Version 2). The majority of the questions were simplified or reworded to improve the agreement. The criteria within the questions were updated or reworded for ten questions. Answer options were updated or reworded where relevant. Sub-domains were updated. The *Value* domain was excluded. While the IRR result for the *Privacy/Security/Ethics/Legal* domain was problematic, the importance of the *Privacy/Security/Ethics/Legal* domain was one of the key findings from the literature review and the focus group discussions. However, we found it is the hardest and the weakest area of the framework to evaluate. Therefore, we assessed this domain separately without incorporating it into the overall scoring, but still keeping it as a domain in the framework. In the final step of Phase II, the framework was finalised with nine domains and 24 questions (Version 3).

Concluding the project: The framework initially synthesised ten domains for the assessment of health apps. These were primarily based on the evidence from our scoping review: *Clarity of Purpose of the App*, *Developer Credibility*, *Content/Information Validity*, *User Experience*, *User Engagement*, *Interoperability*, *Value*s, *Technical Features and Support*, *Privacy/Security/Ethics/Legal*, and *Accessibility*. The framework adopted the most commonly used 5-point Likert scale as a scaling mechanism and the total score for scoring. The average score of the total was used when multiple raters evaluated the same health app.

There are overlaps between some components of different domains, suggesting there is some flexibility within frameworks and that the categorisation of domains is not yet a standardised process. Therefore, we reviewed the ten domains based on our scoping review, the focus group study, and the evaluation process.

Value domain was deemed important based on the scoping review and focus group study. However, this domain, in which perceived value was a sub-domain, was determined to be a highly subjective domain to assess during the evaluation and was therefore removed from the draft framework (Version 3). Some focus group participants had suggested that carefully assessing all the other domains would mean that the "value" of the app has effectively been assessed. Similarly, the *Privacy/Security/Ethics/Legal* domain proved very complex to assess. Therefore, the scoring of this domain was not incorporated into the total scoring for health apps. Moreover, locating information to assess this domain needs thorough training. More research is needed to review the *Value* and *Privacy/Security/Ethics/Legal* domains.

Accessibility was identified as a new domain deemed crucial for assessment due to ethical and equity considerations. The benefit of *Accessibility* as a measurable domain arose in consideration of the framework on COVID-19 apps, where public criticism of the apps highlighted their poor support for CALD communities. In addition, the importance of *Accessibility* was highlighted in the focus group discussions. Our study found that in the mental health and wellness area, it is imperative for apps to support multiple languages and have simple user interfaces that have been ethnographically tested to ensure usability in a mixture of communities and people with disabilities.

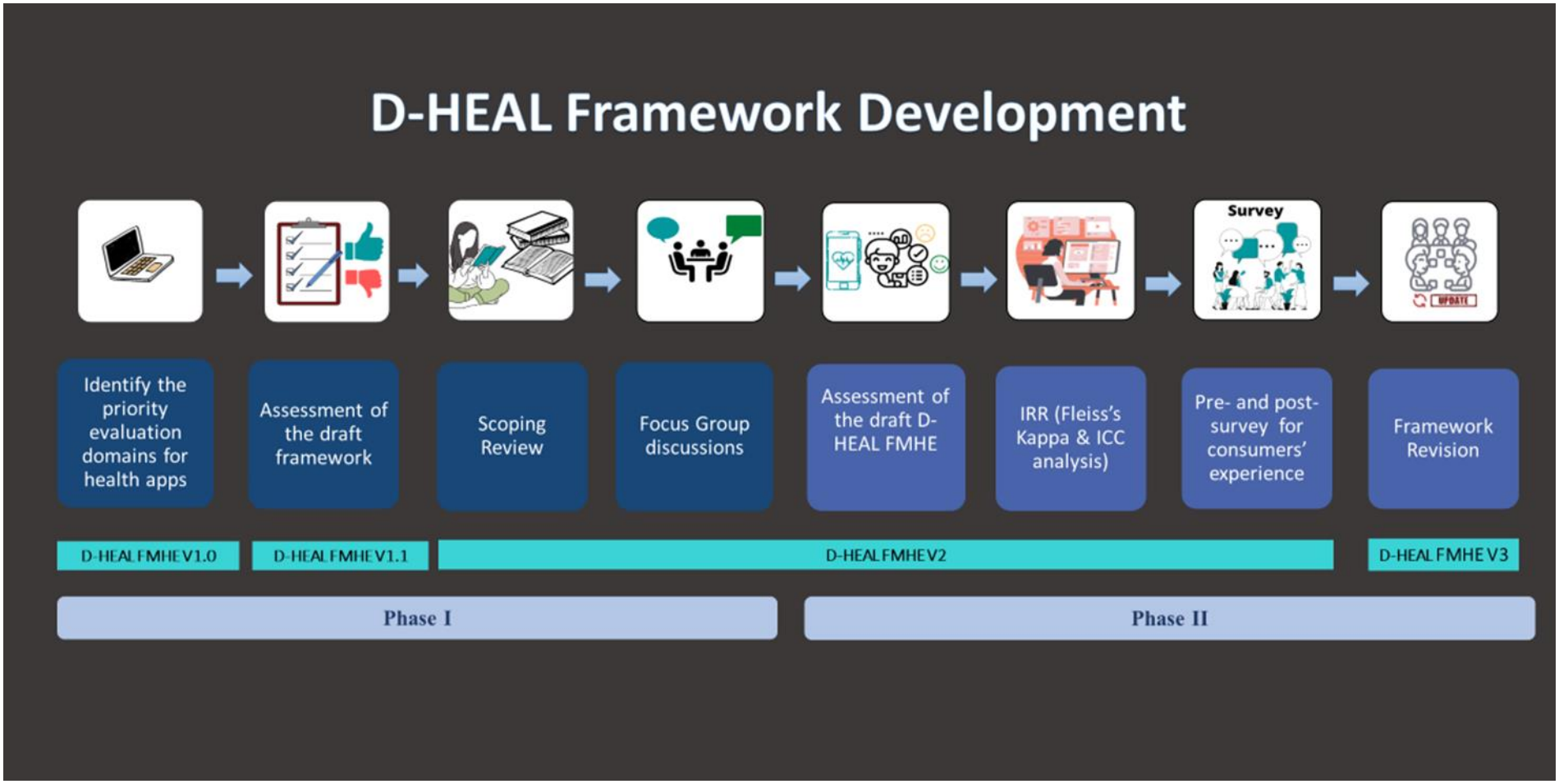
Such aspects of fairness and equity are not frequently considered in other app evaluation frameworks, nor are they a feature of many major apps in the marketplace. Our report highlights the need for the requirement by the app stores to require clear evidence of accessibility support for apps specific to the health domain.

Finally, the D-HEAL framework was developed for organisations to evaluate health apps to promote the best health apps for their clients. Our iterative development and testing process has generated a framework which is suitable for use by organisations, and which provides ratings and recommendations which are consistent with the actual experiences of app users, yet which adds a level of technical assessment that most users will not be able to undertake themselves.

References:

1. Hensher, M., et al., *Scoping review: Development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers.* Journal of the American Medical Informatics Association, 2021. **28**(6): p. 1318-1329.

Pictorial presentation of the D-HEAL FMHE development



Domains


Content/Information Validity


Value


Clarity of Purpose of the App


Privacy/Security/Ethical/Legal


User Experience


Developer Credibility


Functionality

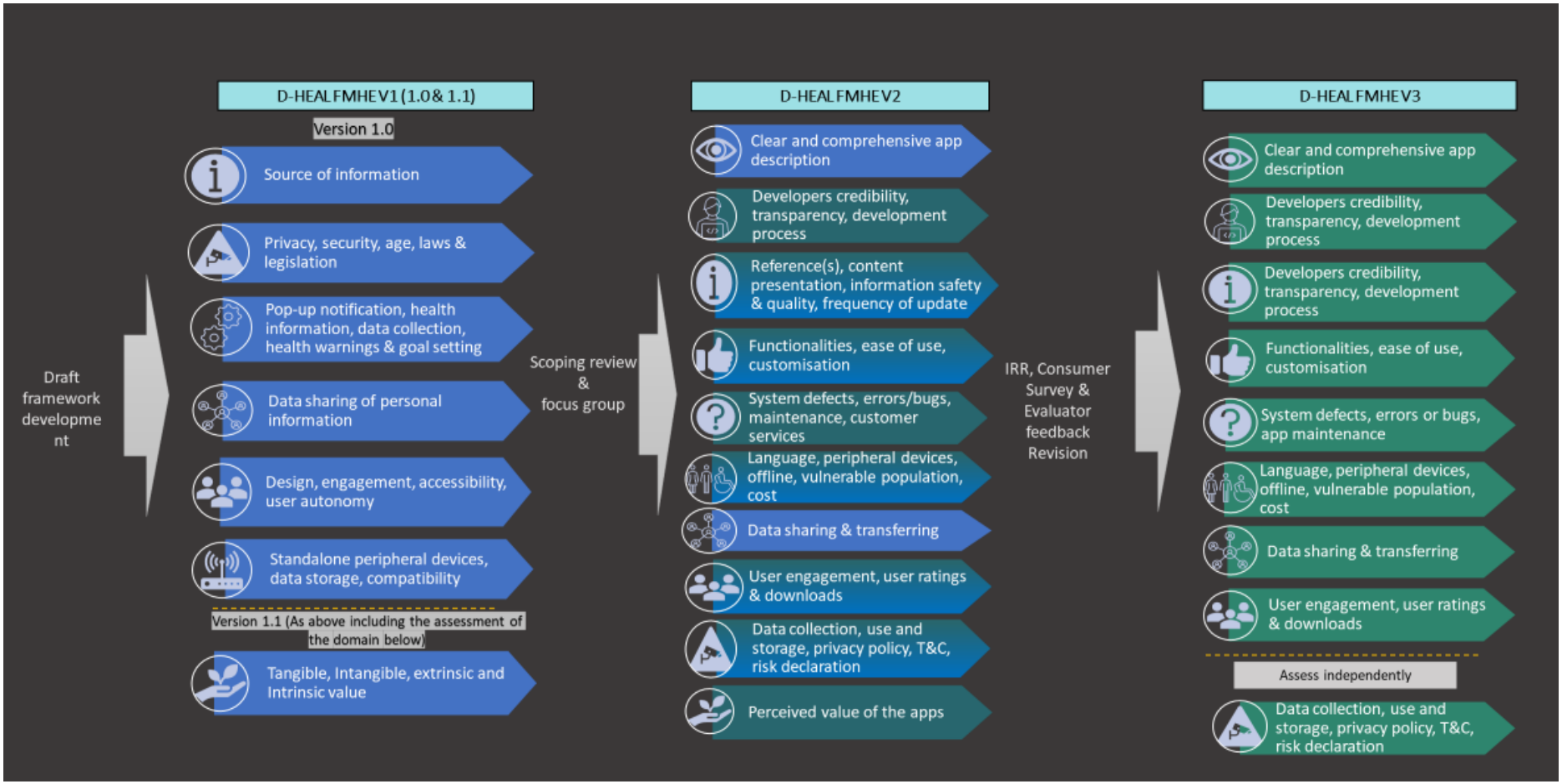

Technology Requirements


Technical Features and Support


User Engagement


Interoperability


Accessibility


Draft framework development
D-HEALFMHEV1 (1.0& 1.1)
Version 1.0
Source of information
Privacy, security, age, laws & legislation
Pop-up notification, health information, data collection, health warnings & goal setting
Data sharing of personal information
Design, engagement, accessibility, user autonomy
Standalone peripheral devices, data storage, compatibility
Version 1.1 (As above including the assessment of the domain below)
Tangible, Intangible, extrinsic and Intrinsic value
Scoping review & focus group
D-HEALFMHEV2
Clear and comprehensive app description
Developers credibility, transparency, development process
Reference(s), content presentation, information safety & quality, frequency of update
Functionalities, ease of use, customisation
System defects, errors/bugs, maintenance, customer services
Language, peripheral devices, offline, vulnerable population, cost
Data sharing & transferring
User engagement, user ratings & downloads
Data collection, use and storage, privacy policy, T&C, risk declaration
Perceived value of the apps
IRR, Consumer Survey & Evaluator feedback Revision
D-HEALFMHEV3
Clear and comprehensive app description
Developers credibility, transparency, development process
Developers credibility, transparency, development process
Functionalities, ease of use, customisation
System defects, errors or bugs, app maintenance
Language, peripheral devices, offline, vulnerable population, cost
Data sharing & transferring
User engagement, user ratings & downloads
Assess independently
Data collection, use and storage, privacy policy, T&C, risk declaration