mhabitat

NHS National Institute for Health Research

MindTech Healthcare Technology Co-operative

Framework for the effectiveness evaluation of mobile (mental) health tools

October 2017

Victoria Betton1, Michael Craven2, Bethan Davies2, Jennifer Martin2, Natalie Nelissen1, Alicia Ridout1, Lucy Simons2

1 mHabitat, Leeds and York Partnership Foundation Trust
2 NIHR MindTech Healthcare Technology Co-operative, University of Nottingham

This development of this framework was funded by NHS England (Digital Development Lab project).

We would like to thank Quynh Pham (University of Toronto) for useful discussions.

Contents

Legend

Topics in purple are explained in more detail in the Further reading section of the Appendix

Words in blue are explained in the Glossary at the end of the Appendix

> Title refers to a theme in the Questionnaire section of the Appendix

Why should I be doing this?

If you are developing, commissioning or reviewing **mobile health** tools (apps and websites for health and wellbeing), this framework is here to make your job easier. The primary intended audience is developers of mobile health tools, but evaluators and commissioners may also find it helpful.

As of 2016, there are over 259,000 health apps on the commercial app stores alone[1], and this number is growing fast. Out of this, we estimate that less than 0.1% have actually published any data that support their claims of being safe and beneficial. Dr Spence (a GP in Glasgow) sums it up quite well: 'health apps are perhaps mostly harmless and likely useless'[2].

The burden to collect evidence for the safety and benefits of digital tools currently lies with the developers, especially in the early stages of development and deployment. In April 2017, an **NHS accreditation process** was launched, which reviews a self-reported questionnaire filled out by the developer[3]. One part of the Digital Assessment Questionnaire asks about available evidence for benefits, termed **clinical effectiveness**. This measures whether the tool provides a benefit to health or wellbeing. This framework helps you answer those questions (if you're a developer) or evaluate the response to them (if you are an evaluator, for example, a commissioner).

The advantages of NHS accreditation have yet to be established However, we expect that accreditation will facilitate commissioning, ultimately resulting in increased use of the digital tool by NHS services. Even if you are not currently interested in accreditation, there are still several important reasons to collect evidence about the safety and benefits of your tool.

Advantages for developers

Firstly, you need to know that the product you have developed does what it is designed to do. Assuming you wish to attract new users but do not yet have a satisfied customer base to provide recommendations about your tool, advertising is important. However, you cannot make strong claims about the benefits of your tool without appropriate evidence, since this is deceptive advertising punishable by law[4]. A high profile case is that of Lumosity, a paid-for suite of 'protective' brain games, which paid $2 million to settle a court case with the Federal Trade Commission after claiming that these brain games resulted in better performance at work/school[5].

Secondly, although most digital tools for supporting health and well-being don't need to meet any legislation (**check if your tool is a medical device**), the public still needs assurances that they are of good quality and are fit for purpose. Due to the large choice of digital tools, it is easy for potential users to go elsewhere if they are not convinced of the value of your product. You probably know that continued **engagement** with digital tools is very low – most people who start using your tool may give up quite quickly. A powerful way to secure user satisfaction and continued engagement is to have an appealing, easy to use tool that produces a benefit that is obvious to customers. For example, this benefit could be the customer tracking their weight in a diet app. Satisfied users are also more likely to write positive reviews and otherwise recommend your tool, attracting more new users. In today's world, most people seek further information before deciding on a product, most often following recommendations from friends or reading reviews online.

1 http://research2guidance.com/product/mhealth-app-developer-economics-2016/
2 http://www.bmj.com/content/350/bmj.h1887
3 https://developer.nhs.uk/digital-tools/daq/
4 https://www.gov.uk/marketing-advertising-law
5 https://www.ftc.gov/news-events/press-releases/2016/01/lumosity-pay-2-million-settle-ftc-deceptive-advertising-charges

Finally, healthcare professionals may want to recommend digital tools to their patients, but need to know which ones are supported by evidence. Evidence-based practice is important to clinicians, and forms a key part of their training. They would not prescribe a drug or therapy without seeing some convincing evidence that it works (and is safe) first, so why would they prescribe a digital tool without evidence? Digital health is still very new to many clinicians so the provision of evidence should help them feel more comfortable and more likely to prescribe or recommend.

Overall, collecting evidence will increase the production costs of the digital tool. The least expensive option involves building in evidence collection from the start (as recommended in this framework). In any case, the above benefits of evidence generation must be weighed up against these initial costs. If done properly, the return on investment will be worth it, especially if you are planning to build a tool that can be scaled up and/or used for a longer time.

An example of a similar cost-effective strategy is the involvement of future users in the design of a digital tool (called **co-design, user-led design, human-centred design** and so on)[6]. While this is more expensive to build into the production process, it dramatically increases the likelihood of excellent user experience and high user engagement and satisfaction as a result of working with target end users throughout the development and design process.

Advantage for evaluators

While it is not clear how NHS accreditation will work exactly, it may facilitate commissioning since the digital tool has been reviewed and approved at a national level.

At this point, accreditation seems to be a binary 'pass or fail' process, and it is not clear how high the threshold is for quality and quantity of evidence. For example, is a single peer-reviewed publication sufficient? How large does the study group have to be? Is qualitative evidence sufficient?

Currently, the academic world is still deciding on which kind of evidence and which study types are appropriate for evaluating mobile health technologies. As a result, regulations and guidelines do not want to commit to setting a strict cut-off without sufficient evidence that this cut-off is reasonable and generally accepted.

Even when the cut-off is clear, you still may want to do your own evaluation in specific situations, for example to choose between two alternative tools or check feasibility of implementing the tool in your local setting. Also, there is a big gap between what should be done and what can be done. We cannot expect all developers to become scientists overnight, or being able to double their budget in order to bring in the necessary resources and expertise.

This framework is written primarily for the developer, but it will show the evaluator what we think is feasible and reasonable to expect of developers.

6 ISO 9241-210: https://www.iso.org/standard/52075.html

How to use this framework?

This framework is primarily aimed at developers, but also provides useful insights for evaluators (for example commissioners). It is based on **other frameworks** as well as our own experience with evaluation of digital tools.

You do not need any prior experience in order to use this framework, though we do recommend **finding partnerships to bring in additional expertise** where possible. This framework will allow you to do simple evidence generation yourself and helps you understand your partners who are generating evidence if you are outsourcing it to them.

We have tried to avoid jargon (when not possible we highlight **words that will link to glossary** and explain them). We have also tried to limit the amount of information you need to read through while still providing you with further reading when you want it. If you don't know much about a topic, you can click **words that link to a section with further reading**.

Each topic is made more concrete by questions you can ask yourself to help guide your thinking, as well as example answers. We have also linked the relevant questions from the clinical effectiveness section of the Digital Assessment Questionnaire, in case you want to apply.

This framework focuses on mobile health tools (apps and websites) in the early stages of evaluation. It will be of most help if used from the very beginning of planning and then throughout development, but it can also be used to evaluate tools that already exist.

Our framework focuses on how you can capture early-stage evidence for the health and/or wellbeing benefits, or the promise thereof, of your tool ('**formative evaluation**'). The aim is to provide insights into the use of a tool to shape future development and testing whilst taking into consideration the time and resource constraints that most developers face.

We also introduce methods for evaluating clinical and cost-effectiveness, which will be required for more formal evaluations of more mature tools, for example by the National Institute for Health and Care Excellence (NICE[1]).

Finally, an important note about legislation. **If your digital tool is classified as a medical device**, you can still use this framework but be advised that you also need to follow the Medicines and Healthcare products Regulatory Agency (MHRA[2]) regulations.

For an overview and summary of the various steps in the life cycle of a digital tool and how this fits in with MHRA, NICE and NHS strategies, you can read **the evaluation life cycle**.

1 https://www.nice.org.uk/
2 https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency

Providing evidence for the benefits of a tool

This section is written for developers of new tools who are at the planning stage. If your tool is at a later stage of development we recommend that you work your way through each of the steps until you reach your current level of development. We will be using questions to help you think about each topic and an example of a fictional digital tool to illustrate this (see excel table).

Idea / planning phase

Before starting a project **it is important** to define the aims and scope of the tool:

- What does it do
- Who is it for
- How should it be used
- What are its anticipated benefits.

Ideally, as much of this as possible should be based on available evidence (**learn more about what is considered evidence**), in order to limit the risk that the final product won't be effective. In some cases, the idea for the tool may be completely new or radically different, so there may not be any direct evidence to support the idea. In that case, try looking for evidence for at least some components of your tool (for example, maybe your intervention itself is new but there is evidence that elements of it, such as peer support and diary keeping, are helpful).

Some of the above may change as you are developing the tool, for example this may be based on suggestions from future users in your co-design workshops, found a new benefit, removed something that didn't work, couldn't reach the original target population and so on. In case of large changes, we recommend revisiting this stage to check your product is still likely to succeed and within budget.

The following questions can help you refine your idea and gather supporting evidence for its demand (will people use it?) and benefits (will it work?). Each question is linked to more in depth questions and the relevant questions of the DAQ, as well as example answers.

> Will there be a demand for your tool?

This section asks you which problem or unmet needs your product addresses and helps you dig a little deeper to see the tool's potential impact and find the target user group.

> Are there indications that your idea/tool is going to work?

This section asks for any relevant existing theory, current practice and information about any offline and online tools/interventions that are similar to the tool that you are going to develop (**how do I look for evidence?**). It is important to decide if each piece of evidence is strong or not (**hierarchy of evidence**), rather than just list them. Strong existing evidence boosts confidence in your digital tool, whereas completely novel ideas may be looked at with suspicion. If you are going for a novel idea, be prepared to generate more evidence for your tool's benefit to convince people that your digital tool is useful.

You may have heard about '**gamification**', the application of gameplay elements (for example points and leaderboards) to encourage **engagement** with a product or service. There are underlying recognised scientific theories about why these work, the so-called **behaviour change**

techniques. If applicable, you will want to cite these as they are considered good predictors for changing a user's behaviour and keeping them engaged.

> How does your tool solve the unmet need?

Here you are trying to break down how your tool will work, which will give you important insights into preparing for the next stage such as which resources you will need, who is going to be your target group for co-design and testing, and how and when you are going to measure 'success'.

If you're new to this, we recommend **using a logic model**, a graphical depiction of the logical relationships between resources, activities, outputs and outcomes. At this stage, you will create a preliminary version, which you will update throughout the next stages as you get a clearer picture of what you're making, what works and what doesn't.

> Updates

Setting out all of this information will put you in an excellent position to start evaluating your tool. However, you should regularly review and update this information throughout development, for example when new evidence is published or when product testing finds that the target population should be changed.

Product development

It is never too early to think about the evaluation of your tool. '**Agile science**', borrowing from principles of agile software, advocates an adaptable and iterative approach to design and development where data is continually collected and fed back into development. Future users can even be encouraged to co-design the evaluation while they are co-designing the look, feel and function of the tool. There are two aspects to gathering evidence:

- What do we measure
- How do we do it.

Evaluating that evidence corresponds to **outcome evaluation and process evaluation** respectively.

When filling out the next set of questions, go back to your logic model and update it where necessary. Maybe during your co-design process you added or removed components, or made something abstract more concrete.

> Instruction manual: how should people use it and why?
> Are people following these instructions and are you reaching the target group?

As a minimum, you should be starting to measure (or putting steps in place to do so) users' usage of the digital tool (how often? for how long do users interact? what do they do? do they drop out early?).

Usage can be tracked with custom code in the tool, or using a free existing services such as Google Analytics or Flurry Analytics (hosted by Yahoo!). Either way, you will still need to include specific code, so put aside some developer time to set this up. Google Analytics reports are for the most part automated or straightforward to set up and can relatively easily be used by non-experts.

Ideally, you do not just want to measure usage, but rather **user engagement**, which includes

whether your users engage in a meaningful way and consider the tool useful and easy to use. Use in a 'meaningful way' reflects whether users are consciously, actively using the tool, rather than just, for example, having opened it accidently (which may still register as 'usage' by analytical software). User engagement is like taking a pill as prescribed; you can't evaluate the medication if people haven't taken it in the correct dose and it's in their body in the right concentration range. Likewise, people will probably need to use the tool in a prescribed way (for example for a specific number of days, how many modules/functions) to be experiencing the anticipated benefits.

It is up to you to define what a meaningful interaction is for your specific tool. This can be based on the evidence you have collected for the potential of your tool (discussed earlier) and/or user testing. For example, if your tool is based on an existing intervention (such as a weight loss programme), how and how often did participants have to behave in the original studies (weigh themselves every week, submit daily calorie intake at least five times a week, read their daily motivational reminder). You may want to explicitly instruct users how to use this tool for optimal benefit.

Exploring user engagement requires asking feedback from your users, for example in the form of questionnaires, one to one interviews or workshops. If done face-to-face or via email/video conferencing, you need to factor in the time and resources to find participants and be able to gain feedback. You should decide how you will contact your users, and whether you have a person to conduct the interviews, and to build and circulate the survey. Alternatively, you can also build evaluation into your tool, for example by having a feedback form or section, or popping up a question to ask users to give feedback. We recommend exploring this possibility in the co-design process, to see what users prefer and find feasible.

As mentioned before, consider the use of **behaviour change techniques** if your user engagement is low or drops off over time. Also, check directly with your users e.g. in a co-design process, what is working for them and what isn't – not all techniques will work for all target groups, or even all people within the same group

> Did you build/test your prototype according to best practice?

DAQ 6 is a related topic: usability and accessibility. It helps ensure that users engage correctly with your product. It asks about user-centred design, which means gaining an understanding of the needs of the correct demographic. This may also be called cooperative design, co-design, participatory design, and human centred design. In addition to getting input from the primary intended users (these may be members of the public, patients, or clinicians, depending on who the digital tool is aimed at), it is helpful to be able to draw on clinical and research expertise where possible.

> Evidence

Finally, you can start collecting evidence for the possible **benefits** of your tool. People may also refer to this as **effectiveness or efficacy**: whether the tool actually works. At this stage, given the fact that you only have a small group of users that you test the tool on for a very short period of time, you probably won't be able to see any real health or wellbeing benefits yet. However, you can start looking at indicators that may predict a future benefit. For example, if your tool aims to help users lose weight, instead of measuring weight itself (your ultimate outcome measure), you can measure predictors such as changes in eating, intentions to change, and actual exercise behaviour.

It can feel overwhelming to try and pick the correct **outcome measure**. There are hundreds of possible scales and questionnaires out there (**see here for a list of the most common outcome measures for children and adolescents related to mental health**). Ideally, you want to pick one that makes sense to the user so they can fill it out on their own. Also, if relevant, you ideally want to use an outcome measure already known and trusted by whoever is going to recommend or commission your tool (for example clinician, teacher, employer). We recommend co-designing your evaluation with users, and – if relevant – the other stakeholders mentioned. Ask them which benefits they expect to get out of using the tool. If there is no existing scale or questionnaire, ask them to design their own; through this they will feel ownership and will be more likely to fill this in when you are testing your tool with them later on. If you make evaluation part of your co-design process, you can possibly save some time and resources because participants are already testing the design and functionality. It does not cost much extra to ask them about benefits at the same time you are asking about ease of use.

How you collect evidence for benefits depends on what you are measuring. It is likely that you will need to ask your users directly (for example: what did you eat today? see methods for obtaining user engagement) but in some instances you may be able to collect data automatically. For example, this could be via activity data measured through a wearable device or phone sensor. In this case, you need to be aware of how accurate the sensor and measurement is.

For acquiring and analysing data about usage, engagement and effectiveness, we give some tips on how to **design surveys** and **analyse data**.

> Possible harms or unintended consequences

Keep in mind that your tool can have unintended consequences. Sometimes these can be positive (for example you've only asked people to exercise more to lose weight but they also report feeling happier), but often there is also risk of possible harms.

You may think your tool has no negative aspects, or 'harms', whatsoever. However, there is always potential for an adverse or harmful outcome to occur. For example, in using a food and exercise tracking app, some people may gain weight instead of losing weight. Or if a person with anorexia picks up your extremely effective weight loss app, it may cause serious harm to their health.

The next topic helps you think about some of these risks. The idea is to list everything that could possibly go wrong as a results of people using your tool, and rate:

- How serious is the adverse effect
- How likely is it that the adverse effect will occur.

These are combined into a clinical risk matrix, such as the example below taken from the NHS Clinical Risk Management in the Deployment and Use of Health IT Systems Implementation Guidance.
Source: NHS Digital[1]

1 http://content.digital.nhs.uk/media/20986/0160382012imp-guide/pdf/0160382012imp-guide.pdf

Likelihood						
	Very High	3	4	4	5	5
	High	2	3	3	4	5
	Medium	2	2	3	3	4
	Low	1	2	2	3	4
	Very Low	1	1	2	2	3
		Minor	Significant	Considerable	Major	Catastrophic
		Severity				

5	Unacceptable level of risk
4	Mandatory elimination of hazards or addition of control measures to reduce risk to an acceptable level
3	Undesirable level of risk Attempts should be made to eliminate the hazards or implement control measures to reduce risk to an acceptable level. Shall only be acceptable when further risk reduction is impractical
2	.Acceptable where cost of further reduction outweighs benefits gained or where further risk reduction is impractical
1	Acceptable, no further action required

Ideally, you should try to find ways to prevent or mitigate any adverse effects which are not green in the above matrix.

If you intend for your tool to be used within the NHS, you will need a formal clinical safety assessment undertaken by a clinical safety officer. We recommend you set up a hazard log where you note risks (potential negative impacts on the user's health or wellbeing) and include mitigations to those risks.

> Possible barriers to adoption of your tool

A related topic to clinical safety, are the factors that could prevent your target audience from using your tool. Some possible reasons why you are not reaching your target audience could be:

- They don't have the right hardware and/or software.
- For example, if your app is iPhone only but your target audience uses mainly Android, or they may not have enough memory on their device to install it
- They don't have sufficient funds to buy your tool, keep up the subscription cost and/or have the data allowance to use your tool continuously.
- For example, children may not have credit to buy apps on the App Store or Google Play and may only have internet connectivity when at home
- They may experience technical problems with the tool itself, including usability.
- For example, slow app with bugs, small low contrast text
- They don't feel capable or motivated.
- For example, brain games for older people only works for those who feel confident to use digital devices
- They don't trust your tool
- For example, it is not clear which information you collect and what you can do with it, they want recommendations from people or organisations they trust.

Ideally, by the end of the development phase, you should have a minimum viable product with approved design, functionality, user engagement and some ideas about possible or potential benefits.

Early product

In this stage, the product is considered sufficiently stable for larger scale deployment outside the initial test group (for example go live on the app stores or internet, or otherwise be tested in a new larger user group). In this stage, we want to answer the question:

Could the digital tool have effects on health and/or wellbeing?

This stage is also called a 'formative evaluation', 'feasibility study', 'pilot' or 'proof of concept'. It means the use of **formal and informal**, **qualitative and quantitative** methods, to explore ways to measure and report outcomes (in this case, the benefits of using the tool). There is no right way to do such an evaluation, as it depends on the tool, existing evidence, expected outcomes and available resources. The aim is to find out what works and what doesn't.

If your digital tool is considered low risk and offers general wellbeing rather than specific health benefits, this stage may be as far as you need to go to help members of the public and clinicians understand the effects on health and/or wellbeing. Otherwise, this stage helps you prepare for the next one, which is summative evaluation - large scale formal trials ideally supervised by someone with a strong research background. Before you start such resource and time intensive trials, you will want to fine-tune your methods and outcome measures. **Read more about what we are working towards as the current but debated golden standard, the RCT.**

At this stage, we are trying to strike a balance between optimising the digital tool and gathering sufficient evidence. Any (big) changes to the digital tool could mean you are changing the way the tool provides its benefits and we may not be able to use the evidence we have collected up to that point (**why we can't pool data**). On the other hand, we do not want to continue testing a digital tool that we know has a serious problem. Therefore, it is important to keep track of any major changes and see if there was an impact on the collected data.

Ideally, you will also want to start looking at whether the observed benefits will outweigh the costs. If you intend to offer your tool to the NHS, you may need to do some form of health economics analysis. This is currently outside the scope of this framework, but be aware that the benefits you are collecting evidence for here will feed into your health economic analysis. **Here's a quick introduction to health economics**.

> Evidence

As mentioned before, there is no right way of doing this. From the previous phase, you should have a fairly good idea about which measures of **user engagement** and benefits to use and how to acquire this data (questionnaire, interview and so on).

We recommend you acquire those measures at least at two time points in the same person: once before the participant starts using your tool (pre-use) and once after a certain period of using the tool (post-use). The second measurement can be after a set amount of time or a certain amount of use (for example if your tool helps manage asthma, some people may get 100 and others only one if you pick a fixed time period). The amount of time/use between pre and post testing depends on how much time/use you expect is needed by users to see any benefits. A suggested general time frame is one to three months.

This type of pre-post repeated measures design offers a good compromise. Compared to having two different groups of people (one of which would use the tool, while the other would do nothing or something different), a repeated measures design offers more statistical power (better chance of picking up small effects, hence smaller number of participants required) but doesn't remove unrelated confounds (for example change over time that would have happened regardless). The

latter requires a separate control group. See next stage regarding control groups.

At this stage, it can still be appropriate to measure mediating factors predictive of future change if primary benefits would be too slow or subtle (like the weight loss example given earlier). It may also be useful to have various post testing sessions, for example at one month, three months and six months, if it is desired to show indications of continued progression. If outcome measurements show quite a bit of variability either due to their nature (for example mood) or the measurement device (for example heart rate via phone sensor), it may be optimal to have several measuring sessions for each time point.

The reason to continue collecting user engagement (or usage if previously validated) at the same time as benefits evidence, is to ensure that people are still engaging correctly. Participants that do not engage correctly can be considered as drop-outs or measurement error, and should ideally be replaced by a new participant. Likewise this can also help explore how engagement changes over time.

There is no rule for how many participants should be included in a study. **Power calculations** offer some insight, but require us to know the effect size. We suggest starting with a sample size of around 40 participants, taking into account not all of these will be useful for data analysis. For example, some participants will drop out, or not complete all questions.

It is advised to continue monitoring and recording any unintended effects, as well as any potential or real harms.

Mature product

For the purposes of evaluation, we consider a product mature if it is in a stable state in appearance and function as far as the user is concerned. It is possible for bugs to be fixed and for software updates to be added periodically to the tool, but no functionality should be introduced or removed, or significantly changed. It is hard to say which changes are significant - therefore ideally it is best to log every change and check if they affect the collected data.

Of course, most products will evolve over time, but the key to evaluation is to test one specific version. Or, in other words, each different version (if it is markedly different from previous version) theoretically needs its own study to prove its effectiveness. Practically speaking, this means you pick a mature version and keep it around for the participants in the study for as long as you need to collect data. Other versions can be made available to the general public in the meanwhile, but your participants should stick with the version they started with. You then try to argue that the new current version still includes the important aspects that made your study a success. If changes were aesthetic and/or extra functionality was added rather than removed, this should be fairly straightforward.

This stage answers the question:

Does the digital tool show ***significant*** effects on health and/or wellbeing?

If your innovation shows promise, it may end up in an independent research study or clinical trial. This is more likely to happen for digital tools that can demonstrate high quality in the first place. We strongly recommend you partner-up with a research team to run and analyse this kind of study, also called summative evaluation.

Summative evaluation or formal assessment is the use of accepted research study designs by the academic community. Or, in other words, the type of design that would be acceptable for

inclusion in a systematic review (see earlier) or clinical trials database. Currently, the **randomised controlled trial (RCT)** is the golden standard for proving effectiveness.

Currently, while many researchers agree that RCTs are not right for mobile health (due to rapid development and obsolescence of mobile technologies and tools), there is no clear consensus alternative. Academic research into alternatives is currently on-going.

Here, we offer some **general advice** but the creation and testing of a new study design adapted to mobile health tools is ***outside the scope of this framework***.

Appendix

Questionnaire

Question numbers, for example Q1, refer to the NHS Digital Development Lab original questionnaire order. DAQ1 refers to the numbering of the Digital Assessment Questions[1].

> Will there be a demand for your tool?

Q2. Describe the problem or unmet need that your digital product addresses

Q3. How do you know this is a problem/need (link to research evidence, work with users etc)?

Q4. Who experiences this problem/need? Be specific as possible, e.g. young people aged 11-16 accessing CAMHS.

Q5. How widespread is the problem/need, does it only affect certain populations/areas?

Q6. What is the impact / are the consequences of the problem/need on health and well-being?

Q7. What is the impact / are the consequences of the problem/need on the NHS?

Q8. What is the impact / are the consequences of the problem/need on society?

Q16. How do you know there is a demand for your product (e.g. focus groups, positive user feedback, high number of active users)?

> Are there indications that your idea/tool is going to work?

Q9. Describe how your digital product solves/addresses the problem/need?

Q10. Which theory/model/approach/assumption(s) (e.g. behaviour change, psychotherapeutic theory) is it based on? Is there evidence for the effectiveness of this theory/model/approach?

> DAQ 1.2.1-2 What type of evidence informed the design and development of the service, how did you find and use them? This evidence base should include as wide a range of published studies and guidelines as possible. There is a perceived hierarchy. If different evidence underpins different elements of the digital service, please indicate which evidence refers to which element.
>
> DAQ 1.2.3. Are behaviour change components that are appropriate for the targeted behaviour incorporated within the service? 'Components' refers to 'behaviour change techniques' and 'behavioural economics elements' that should be incorporated (either as content or format or customer journey) within digital interventions to support behaviour change.'Appropriate' means that there is justification or evidence for the effectiveness of those components to support change of the targeted behaviour.
>
> DAQ 1.2.4 Please provide i) links to literary evidence that support the use of the behaviour change components you have used in your service; ii) a short summary of how this has influenced the development of your digital service;

1 https://developer.nhs.uk/digital-tools/daq/

Q11. Does your digital product digitise a service/tool previously available only in non-digital (e.g. paper) form? If so, which service/tool?

Q12. Why is a digital platform appropriate/preferable?

Q13. What is the treatment or support (digital or non-digital) that would usually be offered to the person with the need?

Q14. How is your digital product (expected to be) different/better than the above treatment/ support?

Q15. If there are any currently available digital products similar to your own, please list them (limit of 3 most relevant digital products) and explain how they differ from your product?

> How does your tool solve the unmet need: who will use your tool and what will be the impact?

Q17. Who will be primarily using your digital product? This can be more than one group. (e.g. the young person)

> DAQ 1.1.3-4 who (characteristics of people) your tool is suitable for and, if relevant, who it is not suitable for (clearly defined in the tool or on its website etc)

Q18. Are there any other people who may interact with your digital product (e.g. caregiver, clinician)?

Q19. Who will benefit from your digital product (can be some/all of the above but can also include people who do not use the digital tool e.g. parents, teachers, GP)?

Q20. What impact will your digital product have on health/wellbeing of the user (primary and other)?

Q21. What impact will your digital product have on health/wellbeing of non-users?

> DAQ 1.1.1-2 health and/or wellbeing purpose of your tool for both members of the public and healthcare professionals (clearly defined in the tool or on its website etc)

Q22. What impact will your digital product have on the NHS? Identify which NHS services (e.g. CAMHS or GP services) and how (e.g. through cost-savings, improved efficiencies, better access, fewer do-not-attends etc).

Q23. What impact will your digital product have on society, e.g. by supporting schools or employers, helping people into work or education?

Q27. What features of your digital tool are designed to help your users meet this goal (monitoring of mood or symptoms, diary, medication reminders, educational content, exercises, social support)?

> Updates

> DAQ 1.2.5-6 documented process to ensure the service is revised in line with any relevant updates to the evidence basis/ new or updated guidelines or systematic reviews.

> Instruction manual: how should people use it and why?

Q28. Is there anything else the user needs to do alongside your digital product to achieve the desired outcome? E.g. use another tool, take drugs, attend therapy.

Q 29. How should people be using your digital product? Describe the ideal / intended user journey.

> DAQ 1.1.5-6 when and how the digital service should be used (instructions for use) in order to gain the health and care benefits claimed (clearly defined in the tool or on its website etc)

Q 30. What is this intended user journey based on, e.g. user experiences/input?

Q31. What do you consider to be a 'meaningful interaction' with your digital product? In other words, what is the optimal experience (attention, interest, affect) and extent of usage (amount, frequency, duration, depth) in order to obtain/expect some benefit?

> DAQ 1.3.1-2 How often should users use the digital health service to realise its benefits? * For how long (less than 60 min, less than 30 days, more than 30 days)

Q32. How has the above been determined (e.g. based on non-digital equivalent, pilot study, similar digital product)?

> Are people following these instructions and are you reaching the target group?

Q34. Which analytics/usage data are you collecting? How are you collecting this data (e.g. Google Analytics, server analytics, self-report through questionnaires or interviews)?

Q35. Are your users currently interacting with your digital product in a meaningful way, as you defined in Question 31? Does the reality of how they use your product match with how you intended for it to be used (are there drop offs or drop outs)?

> DAQ 1.3.3 Data which demonstrates actual usage of the digital service in line with the expected usage pattern and explain how it was obtained. *

Q43. Who is currently using your digital tool (e.g. how many people, demographics, any particular groups)?

> Did you build/test your prototype according to best practise?

Q33. Who was involved in creating your digital product and at which point in the development process? Which approach did you use? Did you co-design with patients, get input/feedback from clinicians etc?

Q44. Do you have a clinical or academic member on your core team? In what capacity do they contribute to your team? (e.g. researcher, analyst, advisor)

Q45. Have you partnered with any clinical and/or academic institutions to conduct formative/ summative evaluations on your digital product?

Q46. Have you done any formative evaluation yet? If so, please summarize the study design and results, if they are available.

Q47. Has your digital product been reviewed and/or endorsed by any individuals/organisations? If so, which ones?

DAQ 6 Usability & Accessibility questions

> Possible barriers to adoption of your tool

Q36. Your digital product is available as a (tick all that apply): iOS native app / Android native app / Web app / Website / Other (please specify)

Q37. Which devices/browsers/operating system versions has it been tested on?

Q38. Is it partially accessible offline? Which components need data connection? If applicable, give an estimate of how much data is typically transferred?

Q39. How do people access your digital product (e.g. through app store)? Can anyone get it or do you need refferal/help?

Q40. How much does your digital product cost to download and/or use. Does it require a one-off fee or a subscription? Do users pay for it or are the costs subsidized, for example by an organisation?

Q41. What are potential barriers to use of your digital tool? Does your target audience have access, is there potential for stigma, digital exclusion etc?

Q42. What steps have you taken (or will you take) to overcome these barriers?

> Possible harms or unintended consequences

Q48. What are the potential harms or unintended consequences of using your digital product? Use may lead to an increase in symptoms (e.g. increases in anxiety as a result of being asked about anxiety) or other negative consequences, such as obsessive use or over-reliance.

DAQ 1.2.4 Please provide iii) a list of potential unintended consequences of your service and how these will be mitigated.

> Evidence

DAQ 1.3.4-5 Are you claiming that there are clinical or behavioural benefits from using the digital service? *What are the claimed clinical or behavioural benefits of the digital service? * Increased adherence to health intervention, easier access to info, easier access to health services, facilitate decision making, save time, save expenses, or improve health outcomes.

DAQ 1.3.6-7 Is there evidence available to demonstrate the claimed clinical or behavioural effectiveness of the digital service? * peer-reviewed publication, evidence published, evidence unpublished, not yet but planned, no. Please provide link(s) to the evidence and provide a short summary of how they supports each of the claimed clinical or behavioural benefits of the digital service.

DAQ 1.3.8 How long will it take for the claimed clinical or behavioural benefits of the digital service to occur? *from <3 months to > 5 years

DAQ 1.3.9-15 Are you claiming that there are economic benefits from using the digital service? cost saving, cost neutral or cost incurring in comparison to standard care? Describe where the savings are likely to occur, costs per user, is there cost and resource impact data available to demonstrate the claimed economic benefits of the digital service? * Published study, health economics model, other data. How long will it take for the claimed economic benefits of the digital service to occur?

DAQ 1.3.16-21 Are you claiming that there are user or societal benefits from using the digital service? Increased efficiency/timeliness, saving costs, enhancing equity, improve quality/safety of services. User benefits for patients and carers could include convenience (fewer appointments, less waiting time, less time off work/school), reduced costs (less travelling) or other efficiencies etc. Societal benefits could include increased community engagement, disability payments avoided, larger workforce, increased productivity, increased earning and consumption, increased tax revenue, increased economic growth etc. How long will it take for benefits to occur

Example case study

While the example app used here is fictional, it is based on ideas and observations in existing apps.

Harmony is an app that wants to help children and young people cope with and prevent self-harm. The idea was born when one of the developers in the team got fed up with the current care for his teenage daughter. She has tried to harm herself repeatedly yet they had to wait six months for a specialist appointment and he feels she doesn't get a lot of quality time with medical staff even now. He notices that some of the things she has to do, such as keeping a diary and rating her mood regularly, would be easier if she could do them on her phone rather than on paper. Also, she has been given a list with self-help activities to try and divert her when the urge to self-harm strikes, but she doesn't carry it with her so doesn't use it much.

Questionnaire answers

Q2. Describe the problem or unmet need that your digital product addresses
Harmony aims to reduce self-harming in teenagers. At the time self-harm occurs, young people are usually on their own, but are very likely to have their phone with them (9 out of 10 children in the UK own a mobile phone according to http://www.childalert.co.uk/article.php?articles_id=322). In addition, few teenagers ever tell anyone about their self-harm, which means they cannot be helped by the healthcare system.

Q3. How do you know this is a problem/need (link to research evidence, work with users etc)?
It is hard to guess how many young people self-harm, since it is underreported. Selfharm.co.uk estimates it could be around 13% of 11-16 year olds. The NSPCC (2015) reported a 14% rise in 3 years in the number of hospital admissions. There has been severe criticism from political and other sources on the inefficiency of the current health care system. (The Guardian)

Q4. Who experiences this problem/need? Be specific as possible, e.g. young people aged 11-16 accessing CAMHS.
The majority of people who report self-harm are aged between 11 and 25. (e.g. depression, eating disorders).

Q5. How widespread is the problem/need, does it only affect certain populations/areas?
Self-harm is more common in girls than boys, and often linked to other mental health conditions. It has been linked to lower socio-economic status, sexual orientation, traumatic life events and substance abuse.

Q6. What is the impact / are the consequences of the problem/need on health and well-being?
Increased risk of further marginalisation, poor mental health, increased mortality and increased risk of suicide for young people. Decreased well-being for family members (stress, worry, possibly resulting in mental health issues).

Q7. What is the impact / are the consequences of the problem/need on the NHS?
Serious self-harm results in Accident and Emergency treatment. Young people who show repeated self-harm will be referred to specialist Child and Adolescent Mental Health Services. In addition to the cost of treatment, there is a great demand and lack of capacity in these services, leading to long waiting times and less serious cases being turned away.

Q8. What is the impact / are the consequences of the problem/need on society?
Social withdrawal due to perceived stigma, decreased performance at work or school, taking time off work/school.

Q16. How do you know there is a demand for your product (e.g. focus groups, positive user feedback, high number of active users)?
We have run a focus group with 15 young people (9-16 years old) who are being treated for self-harm in the NHS. They told us they would value an app as it is discreet and with them 24/7.

Q9. Describe how your digital product solves/addresses the problem/need?
Harmony offers personalised self-help strategies to manage the urge to self-harm (e.g. mood-lifting activities, relaxation exercises and thought challenging activities). It also offers the option to keep a diary, rate your mood and get in touch with professional help (such as emergency phone numbers).

Q10. Which theory/model/approach/assumption(s) (e.g. behaviour change, psychotherapeutic theory) is it based on? Is there evidence for the effectiveness of this theory/model/approach?
The self-help strategies are based on cognitive behaviour therapy (CBT) and dialectical behaviour therapy (DBT). These therapies have been shown to be effective in the treatment of self-harm in young people, though it is still unclear which components/strategies contribute most to their effect (Klonsky and Muehlenkamp, 2007)

Q11. Does your digital product digitise a service/tool previously available only in non-digital form? If so, which service/tool?
Though Harmony digitizes existing strategies from CBT and DBT, the combination as such in an app has not been done before, to the best of our knowledge.

Q12. Why is a digital platform appropriate/preferable?
Reduces stigma (app is discrete and password protected), easily accessible 24/7, and is readily available to most young people

Q13. What is the treatment or support (digital or non-digital) that would usually be offered to the person with the need?
Traditional face to face CBT or DBT provided by specialist child and adolescent mental health services.

Q14. How is your digital product (expected to be) different/better than the above treatment/ support?
Harmony is not meant as a replacement, it is aimed at 1) helping people that don't quite feel ready to seek professional help (overcome stigma, realise that therapy helps), 2) prevention for those that have been turned away due to their condition not being serious enough, and 3) supporting existing CAMHS therapy for people who want support in between face to face

sessions.

Q15. If there are any currently available digital products similar to your own, please list them (limit of 3 most relevant digital products) and explain how they differ from your product?
Calm Harm app, freely available on app stores.

Q17. Who will be primarily using your digital product? This can be more than one group. (e.g. the young person)
The young person experiencing the urge to self-harm.

Q18. Are there any other people who might interact with your digital product, for a example a caregiver or clinician?
Clinicians may want to recommend the app and/or discuss the patient's data (mood scores, diary) as part of their therapy. Currently, the young person would have to show it to them - clinicians do not have access to any data themselves.

Q19. Who will benefit from your digital product (can be some/all of the above but can also include people who do not use the digital tool such as parents, teachers, and GPs)?
Young people who self-harm, their families, teachers, GPs and nurses, charities.

Q20. What impact will your digital product have on health/wellbeing of the user (primary and other)?
Will mainly help the young person manage the urge to self-harm and identify triggers, ultimately reducing the frequency of self-harm and related complications.

Q21. What impact will your digital product have on health/wellbeing of non-users?
Positive effects on health and wellbeing of family members may be more subtle and take longer to manifest.

Q22. What impact will your digital product have on the NHS? Identify which NHS services (e.g. CAMHS or GP services) and how (e.g. through cost-savings, improved efficiencies, better access, fewer do-not-attends etc).
GP and A&E cost savings through reduced emergency/unplanned visits while waiting for treatment, reduced time in CAMHS treatment.

Q23. What impact will your digital product have on society, e.g. by supporting schools or employers, helping people into work or education?
In the longer term a better self-harm management could improve daily functioning and relationships, improved school/work attendance and performance.

Q27. What features of your digital tool are designed to help your users meet this goal (monitoring of mood or symptoms, diary, medication reminders, educational content, exercises, social support)?
Diary (identify triggers), mood score (track progress, find self-help therapies that work best, find triggers), various self-help techniques to try and personalise.

Q28. Is there anything else the user needs to do alongside your digital product to achieve the desired outcome? For example, using another tool, take drugs, attending therapy.
Possibly attend therapy if already in CAMHS, but the app can work as a stand-alone preventative/ supportive measure.

Q 29. How should people be using your digital product? Describe the ideal / intended user journey.

The app should be used whenever the user feels an urge to self-harm, to try and manage the urge and prevent self-harm. They can choose which components to use, the aim is to try and find those that they feel are most helpful and then keep using these as necessary. Ideally, the user will learn how to use the coping strategies without needing the app after a while, so they will no longer be dependent on the app (compare to use of training wheels when learning to ride a bike).

Q 30. What is this intended user journey based on, e.g. user experiences/input?
Input from our focus group users.

Q31. What do you consider to be a 'meaningful interaction' with your digital product? In other words, what is the optimal experience (attention, interest, effect) and extent of usage (amount, frequency, duration, depth) in order to obtain/expect some benefit?
The frequency of opening the app is determined by the frequency with which the user gets the urge to self-harm (highly variable between users), but once in the app, we expect them to complete at least one coping strategy (if feeling the urge) or diary entry (if fairly relaxed) and mood rating to be considered a meaningful interaction.

Q32. How has the above been determined (e.g. based on non-digital equivalent, pilot study, similar digital product)?
We discussed this with our focus group users. They told us how they would normally use coping strategies (as needed, frequency varied widely from person to person) and asked about which frequency they considered useful and feasible for filling in the diary and mood rating (the most voted for option was once per day).

Q34. Which analytics/usage data are you collecting? How are you collecting this data? Are you using Google Analytics, server analytics, self-report through questionnaires or interviews?
The app has built in analytics which collects when users log in, which modules they use and for how long, as well as any mood ratings. The diary entries are analysed on the client side and information is only being send to us about the length and how long it took to write.

Q35. Are your users currently interacting with your digital product in a meaningful way, as you defined in Question 31? Does the reality of how they use your product match with how you intended for it to be used (e.g. are there drop offs or drop outs)?
Of the 15 young people in our original focus group, 10 are still using the app periodically after 14 months. 6 of these log in at least once a week to complete diary entries or mood ratings.

Q43. Who is currently using your digital tool? (e.g. how many people, demographics, any particular groups)
In addition to the 10 people from the focus group (8 female, 12-16 years old), we have a new group of 83 people aged 11-18 years (62 female) currently participating in a feasibility study. These young people were recruited via advertising online asking for young people who have recently experienced episodes of self-harm. We do not have any other information about them at this point.

Q33. Who was involved in creating your digital product and at which point in the development process? Which approach did you use? (e.g. did you co-design with patients, got input/feedback from clinicians etc.)
We co-designed our app with the self-harm young people focus group, with continuous input from 2 CAMHS clinicians (a psychiatrist and nurse).

Q44. Do you have a clinical or academic member on your core team? In what capacity do they contribute to your team? (e.g. researcher, analyst, advisor)
Psychiatrist (advice, researcher), nurse (advice, analyst).

Q45. Have you partnered with any clinical and/or academic institutions to conduct formative/ summative evaluations on your digital product?
The psychiatrist holds a fellowship at the University of Somewhere and helped us prepare the study protocol for the feasibility study.

Q46. Have you done any formative evaluation yet? If so, please summarize the study design and results, if available.
We are currently running a feasibility study. Preliminary results suggest that most users find our app user-friendly and engaging, and analytics reports show that 85% of people use the app at least once per week.

Q47. Has your digital product been reviewed and/or endorsed by any individuals/organisations? If so, which ones?
Not yet, we are working with PsyberGuide who have assigned an expert to review our app.

Q36. Your digital product is available as a (tick all that apply): iOS native app / Android native app / Web app / Website / Other (please specify)
Native app for iOS and Android.

Q37. Which devices/browsers/operating system versions has it been tested on?
Various mobile phones and tablets, 5 most recent versions of each operating system (iOS and plain Android).

Q38. Is it partially accessible offline, which components need data connection? If applicable, give an estimate of how much data is typically transferred?
All components are available offline. When the device is on wifi, analytics data is send to our server automatically in the background.

Q39. How do people access your digital product (e.g. through app store)? Can anyone get it or do you need refferal/help?
Via PlayStore and AppStore

Q40. How much does your digital product cost to download and/or use? Does it require a one-off fee or subscription? Do users pay for it or are the costs subsidized e.g. by an organisation?
Free for people participating in the focus group and feasibility trial, not accessible for general public at this moment. We will make the app freely available for all after the feasibility study.

Q41. What are potential barriers to use of your digital tool? Does your target audience have access, is there potential for stigma, digital exclusion etc?
Fear of stigma: being seen with the app on your phone or device / digital exclusion: young people from a lower socio-economic background may not have access to mobile devices.
Q42. What steps have you taken (or will you take) to overcome these barriers?
Fear of stigma: we tried to address this by giving the app a generic name and non-descrip logo, so it doesn't look like an app for self-harm. There is a password protection on the app in case people want to use it on a shared device.
Digital exclusion: The penetration of mobile phones in the UK is ever increasing, meaning we can reach the vast majority of our intended audience. If we can free up some capacity of face to face services, this can also help digitally excluded people.

Q48. What are the potential harms or unintended consequences of using your digital product?
Use may lead to an increase in symptoms (e.g. increases in anxiety as a result of being asked about anxiety) or other negative consequences, such as obsessive use or over-reliance. Becoming overly reliant on the app (e.g. panic when phone battery dies), misusing some of the

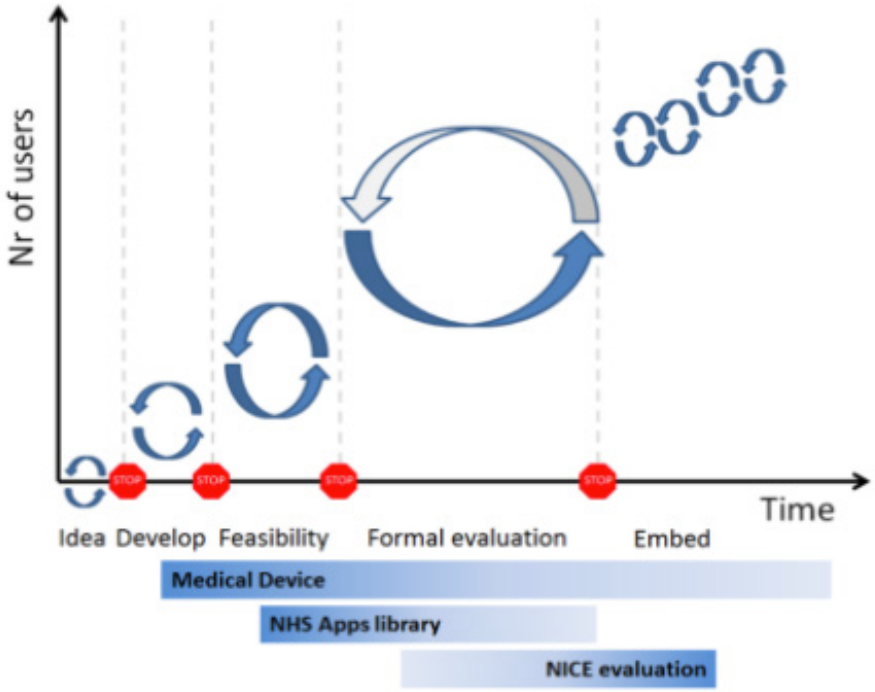
coping strategies, becoming more worried by seeing your diary entries and mood. We have done a clinical safety assessment and mitigated these risks as much as possible.

Further reading

NHS accreditation

In April 2017, NHS England and NHS Digital have launched the new NHS apps library[2] and an accompanying website for developers[3]. The idea is that developers complete a self-assessment questionnaire[4] and if approved, their app will appear on the NHS apps library as 'NHS approved'. The entire process is currently being tested and will become clearer with time. You can read the announcement on this blog[5] on the NHS Digital website.

The evaluation life cycle



Idea	Theoretical groundwork: existing evidence, hypotheses, planning	*Could this work?*
Product development	**Co-design** the product with stakeholders to ensure it addresses the needs and requirements of users Establish **user engagement** and clinical safety (and privacy, technical performance and so on) and keep tracking these throughout product's lifetime Start thinking about **effectiveness** and cost effectiveness	*Do people find it useful and easy to use?* *Do people use it correctly?*
Feasibility (formative evaluation)	Focus on **process evaluation**, qualitative analysis, predictors/ indicators Iterative approach allows methods and digital tool to be adapted, redesigned/edited and re-tested	*Does the tools result in health &/or financial benefits?*

2 https://apps.beta.nhs.uk/
3 https://developer.nhs.uk/apps/
4 https://developer.nhs.uk/daq/
5 https://www.england.nhs.uk/blog/apps-library-is-advance-for-a-digital-nhs/

Formal evaluation	Focus on outcome evaluation, quantitative analysis validated **outcome measures** Improvement over current best alternative (if applicable) Methods and digital tool remain stable **RCT** still golden standard but alternative designs being proposed and tested	*Which health benefits does this tool provide? Is it worth the cost?*
Embed	Keep up to date and implement small improvements while the digital tool is being promoted and becoming 'mainstream'	*Is the digital tool a success in the real-world and sustainable?*

Is my tool a medical device?

This is not an easy question to provide advice on, and in case of doubt, we recommend contacting the MHRA[6].

If your product is software (app or website) that does not come with or interfaces with special hardware (for example wearable device), the most important criterion is whether it has a medical purpose as defined by MHRA[7].

Generally, if your tool makes claims about preventing, diagnosing, monitoring and/or treating a ***specific*** disease, or controlling conception, it may classify as a medical device.

It is not a medical device if your tool is more general, for example providing general advice or recommendations, giving a general risk to a given population (not a specific person), aimed at general wellness of fitness (for example stress or mood vs depression). Also, if your tool replaces existing therapeutic aids (medication reminder, keeping written diary/log) and/or works in conjunction with a clinician (for example app takes picture of skin rash but doctor, not app, decides on outcome), it is likely to be exempt.

It is important to note that MHRA considers explicit and implicit claims you make about the tool through labelling, packaging, promotional literature, advertisements, and so on.

If your tool is a medical device, there are different classes depending on how high a risk it presents. Depending on the risk, you will need to do a self-certification or have a notified body certify your tool. In any case, you will need to set up post market surveillance to track any serious adverse events. Please follow the most current regulations as detailed by MHRA[8].

Which other frameworks are out there?

Since our framework focuses on generating evidence that a digital tool has measurable benefits (effectiveness), we only mention UK relevant guidelines that include at least some advice or standard about this topic. All these frameworks are quite general, and deal with various topics in addition to effectiveness (for example safety, privacy, continuity). There are several frameworks which provide more in depth guidelines about single topics for example user-centred design, information governance and clinical safety, but this is beyond the scope of the current framework. To the best of our knowledge, there is currently no other framework focussing exclusively or in depth on effectiveness.

6 https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency
7 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/610189/Software_flow_chart_Ed_1-03.pdf
8 https://www.gov.uk/government/publications/medical-devices-software-applications-apps

PAS 277:2015[9]
Sponsored by Innovate UK and facilitated by the British Standards Institution (BSI), PAS 277 is a code of practice (a set of recommendations) aimed primarily at app developers. This can be used to help design new tools or evaluate existing ones.

EU working Group on mHealth assessment guidelines[10]
No consensus was reached on an ultimate set of guidelines, but there are interesting discussions on the importance of evidence.

MindTech toolkit for appraising digital mental health products[11]
Offers a common set of criteria for evaluating existing digital mental health tools (apps and mobile websites)

MARS & uMARS[12]
The Mobile App Rating Scale (MARS) was designed by an Australian research team and aims to provide researchers, clinicians and developers with a way to score digital tools based on a list of evaluation criteria.

Why is this important?
Thinking about and writing down your aims and scope will give you an indication about the budget and time you will need as well as the likelihood your tool will be successful. Think about it as doing market research: is there a demand, is there any competition, will people use my product as I intend, will the tool work (i.e. provide a health benefit)?

This allows you to create a business plan, assess your risk of success and do a quick return on investment exercise (are your potential gains going to be higher than your costs). If you intend to distribute your tool in the NHS, be aware that the guidelines are much more strict and you may have to double the budget you need to build the tool in order to ensure compliance with information governance, accessibility, privacy impact assessment, clinical safety and so on.

Even if you are a not-for-profit organisation, you still need to consider drafting a business plan due to ongoing costs such as marketing, hosting and technical support for apps or websites.

In any case, a good initial plan will make all the following phases much easier. You will also need sufficient evidence to convince clinicians (and discerning members of the general public) to recommend or use your tool. The more relevant evidence you can find before building and testing your tool, the lower the burden will be on you to generate new evidence.

The Hierarchy of Evidence
When setting out the idea for your tool it is important to use as much high quality evidence as possible, if it exists, in addition to your own data and results. In healthcare research, there is an established hierarchy of evidence, based on the quality of the research findings.

There are different 'levels' of evidence, often visualised as a pyramid: there is a lot of relatively low quality evidence at the bottom, and as the quality of the evidence increases, the amount of evidence decreases.

9 https://shop.bsigroup.com/forms/PASs/PAS-2772015/
10 https://ec.europa.eu/digital-single-market/en/news/report-working-group-mhealth-assessment-guidelines
11 http://mindtech.org.uk/projects/71-digital-mental-health-tools-evaluation-criteria-technology-theme.html
12 Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps and Simplified user version of the scale (uMARS)

Regulations and guidelines
Theoretically, the highest level should be published regulations and guidelines, which are supposed to be based on all other evidence. However, be aware that anyone can claim to have produced a set of guidelines, so check that the author of the guidelines is an established expert entity (for example BSI or NICE). Also, check if the guidelines are still up to date and relevant (for example to your tool and the countries you want to publish it in).

Systematic reviews
This is a type of literature review that analyses multiple papers/studies using a structured methodology. It often involve a meta-analysis, which is the statistical procedure for combining data from multiple studies. The words systematic review and meta-analysis are sometimes used interchangeably. This type of publications aims to provide a complete summary of current evidence relevant to a research question. Often, individual research studies can produce variable results due to various reasons (for example which methods were used, variability of the measurement, experience of the researchers, engagement of the participants and so on). By combining data from multiple studies, we can remove some of this 'noise' and get a better overall picture. It will give us the bottom line: how strong is the evidence that the intervention work (or doesn't), or is there not enough evidence to decide this just yet? If a systematic review says there is evidence that a certain intervention works, most clinicians and scientists would accept this and be willing to use/recommend the intervention.

Peer-reviewed research papers
This is one study and its results, considered sufficiently rigorous to be published by a panel of usually 2-4 independent scientists (the peer reviewers). Within this category, there is still a wide range of papers, with some seen as more trustworthy depending on the study design (listed below) and the journal they are published in. Certain journals have higher standards and more rigorous review process, while others are considered 'easy' to get published in. One individual study is usually not sufficient to convince a clinician or scientist to use/recommend the intervention, but if there are multiple high quality ones (ideally from different authors) reporting similar results, it is considered almost as good as a systematic review (see note on narrative reviews).

Randomised controlled trial (RCT)
RCTs are considered the golden standard to determine effectiveness (whether the product works) and cost effectiveness (benefits versus cost) of an intervention. Due to their design (for example random allocation, control group, blinding, see section on RCTs) RCTs can determine whether the intervention caused the outcome. Other study

designs can only detect associations between interventions and outcomes at best since they cannot rule out that this association is caused by a third, unknown factor.

Cohort and case control studies

These designs also compare two groups of people, usually to find an association (but not a causal relationship) between risk factors or causes and a disease or health problem.

Cohort studies

In this design, researchers observe a well-defined group of healthy people (cohort) of which some share a common characteristic or experience, for example exposed to a pollutant. The cohort is then followed up to see if there are possible health implications of the pollutant (only an association, not a causal relationship) by comparing the exposed group to the non-exposed group. This can be prospective (study is designed before data collection) or retrospective (using existing data).

Case-control studies

In this design, people with a disease (cases) are selected first and then a matching healthy person is found (control, ideally similar to a patient in all aspects except health status for example age, gender, education, …). The occurrence of the potential risk factor for example pollutant is then calculated and compared between both groups to see if it is higher in the disease group, suggesting it may contribute to development of the disease.

Case studies

These are reports about a single patient, or sometimes a few patients with a similar disease or health problem. Often these are rare cases and become important when larger groups (for example for cohort or case control studies) would take a prohibitively long time to collect.

Best practise and expert opinion

Opinion piece

By this we mean publications by experts expressing their opinion rather than reporting data. Examples are editorials (where one or a few experts give their opinion on a topic or piece of published research) and journal clubs (usually a discussion by junior scientists for example PhD students).

Current best practise

Best practice is normally based on the above evidence, but for new problems or changing circumstances, it is possible that there is little published evidence (yet). Current practice may then be based on what works in the field, and is an agreement of opinions between several clinicians and/or scientists.

Other

This includes non-peer-reviewed or unpublished data, anecdotal evidence (for example patient stories) or opinions of individual clinicians or scientists.

A note on narrative reviews

By narrative reviews we mean any reviews that are not systematic reviews. They do not follow a well-defined methodology. They are usually a descriptive summary of the field only (no meta-analysis) and don't use a well-defined systematic search to find available evidence to include (author may pick sources because they were easily accessible or they fit the author's own ideas).

Narrative reviews are more common than systematic reviews. If there is no systematic review yet, they provide a good starting point for non-experts trying to understand the field. They don't fit easily into the pyramid because their quality depends on the included studies, and to some degree the skill of the reviewer to provide an accurate summary.

How do I look for evidence?

How do I find the right publications?

There are several search engines to find biomedical research papers, including PubMed[13] and Scopus[14]. In our experience, Google Scholar[15] also works well and you can use the same advanced search options (for example look for combinations of words) as in Google. Depending on the field you are researching, you may have to tailor your search.

If you are looking for a popular topic, you may need to try combinations of words and be quite specific to filter out non-relevant publications. You may also want to include the term 'review' to see if there are any reviews on the topic; these provide a good starting place. In general, if given a lot of choice, you can impose limits on the year of publication to narrow down the list to the more recent papers and reviews.

If you are researching for more of a niche topic, you may not find reviews and you may need to try different or related terms. You can also do a normal google search for non-academic publications.

How do I get access to publications?

Depending on the publisher, not all peer-reviewed publications are freely available. If you have a collaborator within academia of the NHS, they may be able to access the paper through their institutional subscription or library. More and more, authors put an early version of their paper (same content, just no typesetting by the publisher) online, often on their own website or that of their institution (Google Scholar pick up quite a few of these). You can also email the corresponding author to ask for a copy directly.

How do I read these if I'm not familiar with the field?

Often you can get the most important information about the outcome of the study from the abstract. The Introduction section can be of interest to learn more about the background and find related papers. The Discussion section often discusses different interpretations and potential flaws, which may be useful if you find two papers with seemingly opposite findings.

Behaviour change techniques

As the name implies, a behaviour change technique or method is any process that has the potential to change psychological determinants of behaviour (for example attitude, self-efficacy, habit). A big effort has been made to create a taxonomy[16] to help people implement and replicate behaviour change methods. To make this more concrete, have a look at some of the techniques summarized by Michie et al. (2013, Annals of Behavioral Medicine 46), listed on page 54-76 in the pdf[17].

13 https://www.ncbi.nlm.nih.gov/pubmed/
14 https://www.elsevier.com/solutions/scopus
15 https://scholar.google.co.uk/
16 http://www.bct-taxonomy.com/
17 The Behavior Change Technique Taxonomy (v1) of 93 hierarchically-clusteredtechniques: building an international consensus for the reporting of behavior changeinterventions

To get an idea about their practical use, the most commonly used behaviour change techniques in mobile technology-based health interventions found by a systematic review[18] were:

- Provide feedback on performance
- Goal setting
- Provide information on the consequences of behaviour generally
- Tailoring
- Prompt self-monitoring of behaviour
- Identify barriers to behaviour/problem solving/identify ways of overcoming barriers.

Gamification, the use of typical elements of game playing (for example points and leaderboards) to encourage user engagement, can be used as a tool to implement behavioural change techniques. For example, providing feedback on performance can be done by offering the user points for good/correct performance.

You will need to determine which behaviour you want to change, which are the appropriate behaviour change methods to do so and implement them correctly. A potentially useful series of articles on these topics can be found here

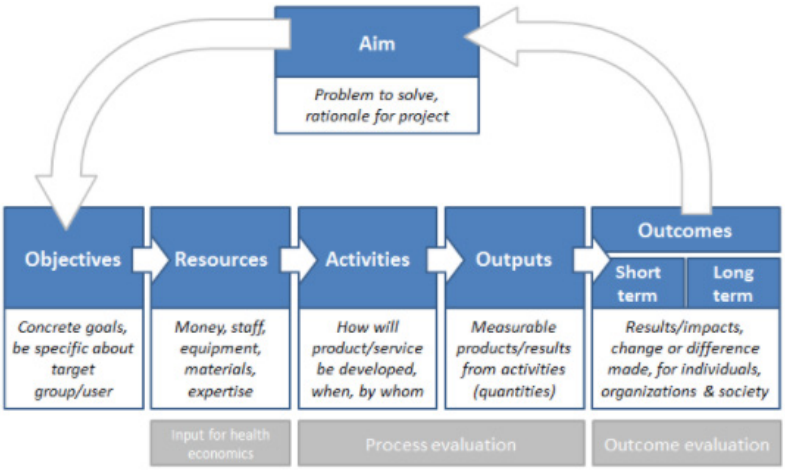
- http://www.ehps.net/ehp/index.php/contents/article/view/ehp.v16.i5.p142/7
- https://www.ehps.net/ehp/index.php/contents/article/viewFile/ehp.v16.i5.p156/8
- http://www.ehps.net/ehp/index.php/contents/article/view/ehp.v16.i5.p190/11

How to create a logic model

A logic model is a commonly used tool to help us think about what we want to accomplish, how we are going to do it and how we will measure success. There are different versions of logic models, using different names and/or components. It's not important which version you use, or even if you create your own adaptation, because it is just a method to help you think about your digital tool.

If you've answered the questions in our questionnaire, you already have the building blocks to start from. The logic model encourages you to explore how they all connect, i.e. find logical relationships between the elements (if A, then B) between them. For example, if I get medication (input), I can then take it as prescribed (output) and as a result my symptoms will reduce (outcome).

Often, this is done in a graphical way, see below. Once you have your logic model, you have a road map for measuring progress and identifying what could go wrong (process evaluation) and for measuring success (outcome evaluation). Outcome evaluation can include effectiveness (outcomes themselves, 'is there a benefit?') and cost-effectiveness (health economics, outcomes combined with resources, 'is the benefit worth the cost?')



18 http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001362

Below is an example of a logic model applied to a fictional weight loss app.

Aim
- App to help user progress to and keep a healthy weight

Objectives
- Encourage healthy eating
- Encourage exercise

Resources
- Developers, graphic designer, UX researcher
- Vouchers for focus group participants
- Social marketing campaign

Activities
- UX researcher uses marketing campaign to find focus group
- UX researcher runs series of workshops with focus group to co-design MVP
- Developers and graphic designer work on creating MVP including
    - Educational component about healthy lifestyles (how to eat and exercise)
    - Food and exercise tracker for users to fill out
    - Motivational messages and reminders
    - Google Analytics to measure app usage
- UX researcher designs questionnaire to measure user satisfaction and weight loss progress
- UX researcher analyses feedback and writes summary report

Outputs
- (At least) 15 participants are included in focus group
- 3 workshops held
- MVP with 4 components
- Change log (changes to MVP as result of workshop feedback)
- 15 questionnaires filled in
- Summary report

(Note that this describes quantity only; the quality of the outputs is assessed by process evaluation)

Outcomes – short term

For the app user:
- Change in buying food and eating out behaviour
- Increase in time spent exercising
- Increased (reported) motivation
- Increased knowledge of how to eat and exercise for a healthy lifestyle

(Note that these are all indicators: signs that predict a possible future weight loss. An actual sustained weight loss takes a longer time to manifest)

Outcomes – long term
- Total weight loss for app user
- Increased self-esteem and confidence for app user
- Positive changes for family (go out more, more active, kids eat better)
- Positive changes for society (increased work attendance and performance, fewer health complications (NHS cost saving))

Some models also include modules like ‘assumptions’ or ‘rationale’, for example you could add the relevant behaviour change methods to each activity.

Other sources:

- Logic model booklet by UCL and Anna Freud
- Innovation Network logic model

Outcome vs process evaluation

According to one scholar, “When the cook tastes the soup, that’s process; when the guests taste the soup, that’s outcome.” In the early stages of development and testing, the focus should be on process evaluation to help you fine-tune your methods, but once you’ve found a method that works, you can go for larger studies and focus on the ultimate aim of proving your tool’s benefits: outcome evaluation.

Outcome evaluation

This form of evaluation focuses on the outcome measures. An outcome measure can be quantitative (for example the result of a satisfaction score between 1-10) or qualitative (for example user comments in an interview such as ‘I like this tool a lot’). Normally, you get several values for a specific outcome measure for example satisfaction scores from different users and/or at different time points. Outcome evaluation tries to analyse these values aiming to summarize the data and draw conclusions from the results. For example, ‘out of 20 users, 90% had a satisfaction score of 8 or higher, suggesting that the tool is generally well liked’ or ‘people who used app A were significantly more satisfied than those who used app B’.

Statistical analysis and thematic analysis are some examples of commonly used analysis methods.

Outcome evaluation will answer the question ‘does this tool provide a real benefit in terms of improved health and/or well-being’.

Process evaluation

This form of evaluation focusses on the process through with the outcome measures are acquired: the methods, implementation and service delivery. Often, it tries to answers questions such as:

- Who did the program reach? E.g. did you reach enough people in your target audience?
- How well was the program delivered? E.g. did your facilitator have the right skills to run the co-design workshops?
- How satisfied were the people involved in the program? E.g. did users feel their comments were taken seriously?

Process evaluation is very useful to see what works and what doesn’t, and to learn from the latter so it can be avoided in the future.

Process evaluation will answer the question ‘which conditions do we need for this tool to work optimally’.

How do I find clinical and/or academic partners?

Bringing in a clinician or scientist with research experience is possibly the easiest way to find existing evidence (literature review and best practise), and design, run, analyse and report on a study.

While some may not be able to fit this into an already busy schedule, a particularly promising group are people working at universities or national institutions. Especially university scientists are required to do some kind of public outreach and prove how their research shows an impact on the world beyond their lab. In addition, they usually supervise a few students who need to do a project, which can be a few weeks to years in case of a PhD. Often, students will work for free as part of their study programme, but still have the prerequisite expertise to be helpful to you. University scientists also have knowledge of possible funding sources and experience with bid writing, and can help you secure funding for larger projects.

We suggest searching your local universities for people working in a field related to your digital tool. Bear in mind that student projects and funding applications take a long time to prepare, so you should start networking as soon as possible. A lot of scientists will welcome the idea of a collaboration provided that their interests align with your project. Their main expectation, especially for larger projects, will be that they can write up the work as a paper for peer-reviewed publication, which helps them secure future funding and you to start building the evidence base for your tool.

Effectiveness or efficacy

Strictly speaking, efficacy refers to how well an intervention works in optimal circumstances (controlled environment, highly trained staff, strict selection of patients and so on). Effectiveness refers to how well it works in the real world and is usually lower (for example people forget to take their pills). Digital tools are generally tested in the real world rather than a controlled environment (you are usually not present when the users test your product), so we can only make statements about your tool's effectiveness.

How to create a survey

First you need to decide which outcomes you want to measure (see logic model) and then figure out the best way to get informative responses. There are a lot of validated questionnaires already out there, check if you find something suitable for your outcomes via Google or Google Scholar. For example, if you want to check if your digital tool improves mental wellbeing, you could use the WEMWBS[19].

You may want to create your own questionnaire for example because there is nothing specifically aimed at your target population (such as children) and/or you want to ask additional questions about indicators, satisfaction and so one. Below are a few tips for creating a survey.

- If possible, create it with input from future users (for example interview a few people to figure out what you may need to ask in order to get the responses you want). Definitely test it on a few people before sending it to your larger study group.
- Open-ended questions (what kind of exercise do you do) and free text responses are useful if you don't know exactly what responses to expect and welcome any feedback. On the other hand, closed questions (do you do high intensity exercise e.g. …) and multiple choice questions are easier to analyse. Adding a multiple-choice option such as 'Other, please specify' can be a good compromise if you're not sure you've listed all available options.
- You can use Likert scales, (often) a five point scale which is used to allow the individual to express how much they agree or disagree with a particular statement for example
    - Strongly disagree / Disagree / Neither agree nor disagree / Agree / Strongly agree
    - Poor / Fair / Good / Very good / Excellent
    - Extremely important / Very important / Quite important / Somewhat important / Unimportant
    - On a scale from 1 to 5 where 1 is casual and 5 is formal: 1 / 2 / 3 /4 / 5

19 http://www2.warwick.ac.uk/fac/med/research/platform/wemwbs/

Tips on analysing data
Often, you will have 2 types of data:

- Qualitative: for example interview responses and free text in questionnaires
- Quantitative: for example scores (such as mood, satisfaction) and percentages (for example % returning users)

There are a variety of ways to analyse data, here we just give a few examples of commonly used methods.

Thematic analysis
The aim is to summarize qualitative data by finding and analysing underlying patterns (themes). It does not depend on any pre-existing theoretical framework (though it can be used with one). Guidelines how to do thematic analysis, including an example, can be found here.

Descriptive statistics
As the name implies, this form of statistics simply describes the data you have. Examples are averages, percentages, maxima and minima. For example, in our study, people who used our app lost on average 5 kg over 6 months whilst the control group lost 1 kg on average.

Inferential statistics
This goes beyond the data you have, and tries to draw conclusions (inferences) about the wider population or general conditions. For example, do people who use the app lose more weight than the control group? Here are some of the commonly used statistical tests:

- One-sample t-test: does the average in your data differ from a known average (for example a population average or '0' if you are testing whether your app has an effect compared to no effect)
- Independent samples t-test: are the averages of 2 (unrelated) groups different, for example does measurement X differ between the intervention group and the control group
- Paired t-test: When comparing 2 measurements in the same group (for example a pre versus post measurement in the same people), you use a paired test for example do people lose weight after using the app compared to their baseline weight
- One-way analysis of variance (ANOVA): are there any differences between the average of three or more (unrelated) groups

Note: If you think you know which way the effect should go (for example expect intervention to do better), you can use a one-tailed test. If the effect could go either way, you should use a two-tailed test.

Doing a statistical test isn't hard (for example Excel can do the most common ones for you), but it is important to be aware that a given test is only valid if its assumptions are met. This is beyond the scope of this framework, but a good tip is to look at your raw data before doing any test: are there any outliers (really big or small values compared to other data points), do the data look skewed? For example it could only have high numbers of low numbers, and nothing in between. Ideally you want your data looking quite continuous, without big gaps or outliers.

When not to pool data
In the same way you can't compare apples and oranges, you cannot make a general statement (summarise the results) if you have different types or quality of data. That's why it's important to keep your digital tool and the way you acquire data as stable as possible for the duration of the study.

For example, if your weight loss tool originally included modules for tracking meals and exercise, you can't suddenly add a forum and expect that people will be just as likely to lose weight (peer support tends to improve weight loss). Or, more subtly, if you add a new function where you provide feedback on what the user is tracking (for example if they do over 10,000 steps, you send them a message and virtual reward), this could motivate or demotivate (if they can't keep it up) users.

An example of a change that is unlikely to affect weight loss but could potentially get you more users is changing the font type or background colour for example to make something easier to read for users with bad vision. However, it is not impossible that there could be subtle effects for example due to effects of colour on mood. Therefore, we recommend logging all changes and checking if there is an effect when you are doing the analysis (for example do users of the red coloured app show more weightless than those who used the grey one).

RCT

The golden standard study design for medicinal drugs and therapies is a randomised controlled trial (RCT). While this is not a feasible design for mobile health tools for various reasons (not in the least being a big commitment in terms of time and funding), we can still try to apply some of its principles to reduce bias.

An RCT study design contains at least two groups (treatment and control) to which people are randomly allocated, the only difference between which is the 'treatment' that is tested, in this case the digital tool. Random allocation avoids selection bias such as person specific factors (for example if given a choice, fit people may pick a fitness app over no app and produce better results), time of year (for example mood effects differ in summer vs. winter), and so on. The control group can be using no app, standard treatment or placebo. While placebo is considered golden standard (to rule out placebo effect), it is not always possible and placebo apps are not trivial to construct (but not impossible, as shown by Sleepio).

We suggest a gradual approach, starting with relatively simple pilot studies and progressing to studies more closely resembling RCTs as the digital tool stabilises and the desired effects are optimised.

A quick introduction to health economics

Once you have measured the effectiveness and benefits of your tool, the next step is to see if this is worth the cost. The idea is similar to other economic calculations you may already be doing, like the return on investment calculations: is the cost of producing and maintaining this product going to be offset by its potential gains.

In health economics, we look at whether the potential benefits are worth the cost of delivering the intervention, either because the benefits outweigh the cost or because the cost to benefit ratio is more favourable than for a competing intervention. Currently, the NHS budget isn't increasing as rapidly as the demand for services, which means that any major investment in a new tool/therapy/drug is likely to replace an existing one. This means there should be convincing evidence that the new tool/therapy/drug is worth more than an existing one.

In health economics, the costs often do not include developing and testing your digital tool. At the starting point, it is assumed you have a functional tool which you are selling to the NHS or Council. The costs and benefits in the health economics analysis are those for the buyer, for example the NHS or Council.

The most common analyses in health economics are cost-benefit analysis and cost-effectiveness analysis. A third form, cost-utility analysis, is a special case of cost-effectiveness analysis. It is used by NICE to determine which interventions are acceptable and may therefore become important for mature digital tools that aim to enter the NHS.

Cost-benefit analysis

This analysis lists all benefits and costs related to an intervention within a specified time period, expressed in monetary terms and discounted (converted to their net present value). Because everything is expressed in monetary terms, there are controversies, for example how to put a value on human health and life, or on the environment.

Cost-effectiveness analysis

Cost-effectiveness analysis avoids monetizing health effects, but it needs a comparison (for example treatment as usual or another control group). The total net cost (including program costs minus the costs averted, expressed in monetary units) of an intervention is calculated and expressed as a ratio to some natural health unit for example cases prevented or lives saved. This natural health unit can also be an intermediate outcome, for example decreased blood pressure, rather than the final outcome, for example heart disease prevented. Importantly, you can only compare such a cost per health unit ratio to interventions impacting the same health outcome (same denominator).

Cost-utility analysis

This analysis is similar to cost-effectiveness, except that the denominator of the cost/health ratio is usually expressed in quality-adjusted life years (QALYs). QALYs are a generic measure of disease burden, taking into account the quality and the quantity of life lived. They are calculated by multiplying the number of years lived in a given health state by the utility value associated with that health state. The utility is usually a value between 1 (perfect health) and 0 (dead), though negative values are possible to reflect states 'worse than dead'. For example

- Living for 1 year in perfect health (utility is 1) equals 1 QALY
- Living for 1 year with a disability with utility 0.5 equals 0.5 QALY
- Living for half a year in perfect health also equals 0.5 QALY

You can see from the last 2 examples that both the quality and quantity of life contributes to the total QALY.

In order to be able to calculate QALYs, and hence do a cost-utility analysis, you need to collect the right data at the time you are also collecting your other outcomes and costs. The easiest way is to use a standard questionnaire such as the EuroQol five dimensions (EQ-5D) questionnaire[20], a standardized instrument for measuring generic health status. The thus measured health states can be converted into utility values based on country specific value sets[21]. You will also need the time period in years for which this health status is true, which may mean following up your participants over time.

General advice about summative study design

Our proposed design comprises two groups of participants, randomly allocated to the digital tool or the appropriate alternative: (no digital tool, standard care or placebo. These groups will be tested before the use of the digital tool (or start of the control condition) and a given amount after using the digital tool (or being in the control condition). Similar to the previous stage, the amount of time between pre and post depends on expected effects, and using more than two time points may be useful to show progressive effects.

20 https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf
21 https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/valuation/choosing-a-value-set/

Such independent group designs tend to require greater numbers of participants due to high inter-subject variability in outcome measures, and hence may take longer to run. They do, however, offer to measure any biases in the control group, which isn't expected to change pre versus post, that would otherwise (as in the previous stage) have been attributed to the digital tool.

For complex digital tools with multiple features, or if there are competing digital tools, more than two groups would be needed. For complex digital tools, it is advised to find out which possible combination of features of the digital tool works best (factorial design; one group for each combination of features).

It is important to match all groups as closely as possible for any potential confounding factors. Most commonly this would be demographics: gender, age, education and so on., disease severity/progression indicators and - specific to digital tools - possibly digital literacy status, but there may be specific factors depending on the digital tool. For example, a weight loss app, factors such as BMI and physical fitness would apply.

If there is a serious competitor digital tool or alternative therapy, this also requires an additional participant group to be added in order to proof that the digital tool is significantly better than all alternatives.

As in the previous stage, user engagement and clinical safety should be monitored, since they can influence the results.

Common outcome measures for young people's mental health

Below are some freely available measures which have been used in previous research evaluating interventions in children and young people. This is not a definitive list but rather a starting point of measures for you to consider using in your own evaluation.

DEPRESSION

Mood and Feelings Questionnaire (MFQ)-Child report: Suitable for use with 8-17 year olds. 33-item questionnaire covers a broad range of depression symptoms in the past two weeks. A shorter version (13 items) and parent self-report version (completed by parent) are also available. http://devepi.duhs.duke.edu/mfq.html

Patient Health Questionnaire – 9 item version (PHQ-9): Popular measure often used with adult community populations (including in primary care), but has been used with adolescents and young people (aged 13-25 years) as well. It consists of nine items assessing occurrence of depressive symptoms in the previous two weeks. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/

Centre for Epidemiological Studies- Depression Scale (CES-D), modified version for Children and adolescents (CES-DC): The CES-D is for adult populations (aged >18 years) and the CES-DC is for children and adolescents aged 6-17 years. It consists of 20 items and looks at depression symptom occurrence within the past week. http://www.psych.uic.edu/csp/images/stories/physicians/rating%20scales/CES-DC.pdf

Revised Child Anxiety and Depression Scale (RCADS): Suitable for use with 8-18 year olds. Consists of 47 items and includes several subscales which assess social phobia, panic disorder, major depression, separation anxiety, generalised anxiety, and obsessive-compulsive disorder. A short version is also available.
http://www.childfirst.ucla.edu/Resources.html

ANXIETY

Spence Children's Anxiety Scale (SCAS): 44-item self-report questionnaire which assesses specific anxiety symptoms relating to six subscales (generalised anxiety, social phobia, separation anxiety, panic attack/agoraphobia, obsessive-compulsive disorder and physical injury fears). There is also a parent self-report version of the measure.
http://www.scaswebsite.com/

Screen for Child Anxiety Related Disorders (SCARED): 41 item child and parent self-report measure used to screen for childhood anxiety disorders including general anxiety disorder, separation anxiety disorder, panic disorder, social phobia, and school phobia (each can be scored separately). Suitable for ages 8-18. http://www.pediatricbipolar.pitt.edu/content.asp?id=2333

Preschool Anxiety Scale-Revised (PASR): 28 items, completed by parents for assessing anxiety in children aged 6 or younger. Four subscales assess generalised anxiety, social anxiety, separation anxiety, and specific fears.
http://www.scaswebsite.com/1_5_.html
A teacher-rated version of the scale has also been developed: http://www.scaswebsite.com/index.php?p=1_65

Generalised Anxiety Disorder Scale (GAD-7): this popular measure is often used with adult populations (over 18 years) but has also been used in adolescents aged 14-16 years (e.g. Wong et al. 2014 - www.sciencedirect.com/science/article/pii/S221478291400013X).
http://carybehavioralhealth.com/wp-content/uploads/2011/06/Generalized-Anxiety-Scale.pdf

Children's Automatic Thoughts Scale (CATS): 40 item self-report measure to assess negative self-statements over the previous week, which can be important in the development and prolonging of anxiety. Designed for children and adolescents aged 8-17 yrs. Four separate subscales of cognitive content are assessed: physical threat, social threat, personal failure, and hostility.
http://www.mq.edu.au/research/research-centres-groups-and-facilities/healthy-people/centres/centre-for-emotional-health-ceh/resources/child-and-adolescent-questionnaires

Penn State Worry Questionnaire for Children (PSWQ-C): One of the most frequently used instruments to assess worry in children. It is 14 item self-report measure, suitable for 7-17 years.
http://www.childfirst.ucla.edu/Resources.html

SELF-HARM

Inventory of Statements about Self-Injury (ISAS): This inventory has been used with young people (aged 16+) and consists of two sections. Section 1 looks at lifetime frequency of self-injurious behaviours (e.g. history of self-injury, type of self-injury). If young people endorse more than one of the self-harm behaviours in Section 1, they proceed to the next section. Section 2 asks about functions of self- injury (e.g. feelings associated with self-injury) and consists of 39 items categorised into 13 functions.
https://www2.psych.ubc.ca/~klonsky/publications/ISASmeasure.pdf

TRAUMA

Revised Post Traumatic Growth Inventory for Children (PTGI-C-R): Assessment of post-traumatic growth or change resulting from a response to trauma. Consists of 10 items and reflects five key domains of growth (two items for each domain): new possibilities, relating to others, personal strength, spiritual change and appreciation of life

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2827205/

KINDLR measure: 24 item generic instrument for assessing health-related quality of life in children and adolescents aged 3 years and older.
https://www.kindl.org/english/information/

When Bad Things Happen Scale (WBTH): 95-item self-report inventory which measures post-traumatic stress disorder symptoms for children aged 7-14 yrs. Contact author for permission.
https://www.ptsd.va.gov/professional/assessment/child/bad-things-happen.asp

DISTRESS

Depression, Anxiety and Stress Scale (DASS-42): 42 items categorised into three subscales measuring depression, anxiety and stress. Scores are tabulated as a whole and for each subscale. A shorter version of 21 items is also used (DASS-21). The scale is often used in adult populations and may not be suitable for young people aged 17 and younger.
http://www2.psy.unsw.edu.au/dass/

Kessler Psychological Distress Scale: 10 item (K-10) and 6 item (K-6) versions: simple checklist that assess symptoms of anxiety and depression in the past four weeks. These measures have been used in adult populations, with some evidence suggesting that they could be administered with adolescents (e.g. Mewton et al 2016).
https://www.hcp.med.harvard.edu/ncs/k6_scales.php

Brief screening scale for Anxiety and Depression (PHQ-4): derived from the PHQ-9 and GAD-7, this short scale asks four items relating to core criteria for depression and anxiety.
http://www.psychiatrictimes.com/all/editorial/psychiatrictimes/pdfs/scale-PHQ4.pdf

The Adolescent Perceived Events Scale (APES): self-report measure of stressful events that commonly affect adolescents. Consists of 90 possible events, ranging from major life events to daily events. The APES can be scored in a variety of ways, including calculating total weighted scores for both negative and positive events or generating separate scores for major and daily events.
http://vkc.mc.vanderbilt.edu/stressandcoping/apes/

WELL-BEING AND COPING

Warwick-Edinburgh Mental Well-being Scale (WEMWBS): 14 items that measure different aspects of positive mental health over the past two weeks. The scale is validated for use with individuals aged 13 to 74 yrs. There is also a shorter version with 7 items which relate more to functioning than to feelings, offering a slightly different perspective on mental well-being.
http://www.warwick.ac.uk/fac/med/research/platform/wemwbs/

Stirling Children's Wellbeing Scale (SCWS): 15 item self-report measure. Suitable for children aged 8-13 years. There are three subscales assessing positive emotional state, positive outlook, and social desirability.
http://www.warwick.ac.uk/fac/med/research/platform/wemwbs/faq/scwbs_children_report.pdf

Strengths and Difficulties Questionnaire (SDQ): Screens for 2-17 year olds' behaviours, emotions, and relationships. Different versions are available. Each version includes a section of 25 items on psychological attributes divided between 5 scales (emotional problems , conduct problems, hyperactivity/inattention, peer relationships problems and prosocial behaviour). Extended versions ask whether the respondent thinks the young person has a problem and if an

intervention had any impact. There are self-report versions for 11-17 year olds as well as parent/teacher-rated versions.
http://www.sdqinfo.org/py/sdqinfo/b3.py?language=Englishqz(UK)

General Self-Efficacy Scale (GSE): 10 item scale which assesses self-beliefs and perceived confidence to cope with a variety of life demands. Suitable for adolescents and adults aged 13 and older.
http://userpage.fu-berlin.de/~health/engscal.htm

Self-Efficacy Questionnaire for Children (SEQ-C): 24 items about academic, social and emotional self-efficacy (can be used separately). Aimed at secondary school age (14-17 years).
http://drjenna.net/wp-content/uploads/2013/07/self-efficacy_questionnaire.pdf

Rosenberg Self Esteem Scale: 10 statements related to overall positive and negative feelings of self-worth or self-acceptance. The original scale was tested with adolescents, but may not be suitable for children (Hosogi et al 2012).
http://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Self_Measures_for_Self-Esteem_ROSENBERG_SELF-ESTEEM.pdf

The Adolescent Coping Orientation for Problem Experiences Inventory (A-COPE): 54 items describing different behaviours that adolescents (11-18 years) may use to cope in stressful situations.
http://www.performwell.org/index.php/find-surveyassessments/outcomes/emotional-wellbeing/self-management/adolescent-coping-orientation-for-problem-experiences-a-cope#

The Child-Adolescent Perfectionism Scale (CAPS): this 22-item measure assesses perfectionist standards, categorised into two subscales of self-oriented perfectionism (what the young people themselves think about their standards) and socially-prescribed perfectionism (what other people think about the person's standards).
http://hewittlab.sites.olt.ubc.ca/files/2016/03/CHILD-ADOLESCENT-PERFECTIONISM-SCALE.pdf

Children's Global Assessment Scale (CGAS): To be completed by an adult (e.g. parent, clinician). They rate the child's current level of functioning on a scale of 0 (most serious impairment) to 100 (best level of functioning).
http://www.corc.uk.net/outcome-experience-measures/childrens-global-assessment-scale/

Columbia Impairment Scale – Youth Version (CIS): 13 items that assess multiple areas of psychosocial functioning, including interpersonal relationships, occupational or academic functioning, and use of leisure time, in addition to some questions on broad areas of psychopathology (e.g., feeling sad or unhappy). The CIS can be completed by a parent or other adult as well as a child/adolescent.
http://www.dhs.state.il.us/page.aspx?item=86069

Generalised Anxiety Stigma Scale (GASS): This scale measures stigma associated with anxiety disorders. It has two subscales with 10 items each, which measure two different types of stigma: personal stigma (stigma in the respondents own attitudes) and perceived stigma (respondent's perception about the attitudes of others). Need to request permission to use.
http://cmhr.anu.edu.au/mental-health-measures/generalised-anxiety-stigma-scale-gass

Glossary

agile science

A term coined by Dr. Eric Hekler[22] (Arizona State University) which is based loosely on the principles of agile software development. It indicates an adaptable and iterative approach to the design and implementation of evaluation studies. Instead of designing one large study and then running it as planned, agile science advocates smaller rapid cycles of designing and testing, which feed back into a cycle of continuous development.

co-design (also participatory design, user-led design, human-centred design)

The term co-design comes from co-operative design, which refers to involving all stakeholders in the creative process of formulating and solving a problem, in this case create a new digital tool. Stakeholders are usually the users of the tool, but can also include other people with knowledge of or a vested interest in the tool such as family, commissioners, clinicians, scientists, and so on. Co-design helps create a product that is relevant and appropriate, improving user satisfaction and engagement.

effectiveness

Effectiveness means the degree to which objectives or intended outcomes are achieved. In our framework, we mean the specific instance of clinical effectiveness, i.e. realised benefits for health and/or wellbeing.
See also **effectiveness or efficacy**

(user) engagement

This is the user's response to the digital tool: do they use it in a meaningful way, do they find it useful and easy to use. The definition of meaningful interactions depends on your specific tool (for example fill in food diary at least once a day versus go on the app when feeling an urge to self-harm and complete at least once coping technique). An analogy is taking a drug according to prescription. If users don't get the right dose of your tool at the right time, it is unlikely there will be any effect on their health or wellbeing.

formal versus informal

By 'formal' we mean something (for example a procedure or study) which is generally recognised as being a 'correct' way given the situation, according to expert opinion. This is typically done in an academic or professional setting, follows a set of rules/recommendations, is more complicated and takes longer.
For example, a randomised control trial is a formal way to test the effectiveness of an intervention.

formative (evaluation)

This is a term often used to contrast with summative evaluation.
Formative means it helps shape the process/tool and is used in the first stages of testing a tool (development and early implementation). The primary intended users are the staff testing the tool. It suggests small and incremental modifications to potentially improve the process/tool.
Summative (formal) refers to determining effectiveness/efficacy and is used for a mature product. The primary intended user is usually the person who commissioned/funded the evaluation. It provides suggestions whether to continue, stop or (substantially) modify the product.
See also **effectiveness or efficacy**

gamification

The application of game design elements and principles to encourage user **engagement** in a non-game context, in this case mobile health. Some popular elements are setting challenges and tracking progress (levelling up), rewards (for example points and badges), competition (for

22 https://www.ncbi.nlm.nih.gov/pubmed/27357001

example leader boards) and adding narrative (for example fulfil an epic quest).

mobile health (mHealth)
The use of mobile (wireless) devices (mobile phones, tablet, wearable devices for example fitness tracker) supporting medical and public health practice.

outcome measure
A test used to measure the impact of the digital tool. This can be a performance test (for example weight or step count) or self-reported measure (for example mood rating or user satisfaction). Outcome measures are often created by experts, for example scientists and clinicians, though there is a growing trend to also involve users in order to find out what is meaningful to them (for example patient centred outcome measures). You can find existing outcome measures by **doing a literature search**. Often, such outcome measures have been validated, i.e. there is evidence that they are accurate and acceptable (for the described use). You can also create your own outcome measures, possibly co-designed with the users of your digital tool, but bear in mind that these will not be as widely accepted by experts (for example funders) as validated outcome measures.
See here for a list of the most common outcome measures for children and adolescents related to mental health

process evaluation
This is a term often used to contract with outcome evaluation, see also **outcome vs process evaluation**

power calculations
In statistics, the power of a test is the probability of detecting an effect, if it truly exists. An example of an effect could be weight loss for a dieting intervention. If you know the size of the effect (for example from existing literature or a pilot study) then you can calculate the minimum sample size (for example number of digital tool users) you will need to be able to detect it. Or vice, versa, if you know the sample size, you can calculate the minimum effect size you will be able to detect.

qualitative versus quantitative
Quantitative data are about a quantity, i.e. can be expressed as numbers (for example weight, user satisfaction on a score from 1 to 10) and can be analysed using statistics. Qualitative data describe a quality, something that can be expressed as words or images but not numbers (for example the experience a user had in their own words). Qualitative data are harder to analyse and more subjective. Qualitative research is mainly exploratory and can be used to gain more insight into a problem (for example find underlying motivations of why people think or behave in a certain way) and suggest hypotheses for quantitative research (for example quantify the motivation and make statements about people in general rather than specific individuals).

randomised controlled trial (RCT)
This is currently the golden standard study design for clinical trials testing the **efficacy or effectiveness** of medicinal drugs and therapies. Participants are randomly allocated to the treatment group or a control group not undergoing active treatment. The randomisation and control group help reduce bias (for example person specific factors, time of year).
For more details, see **RCT**.