COULD MEDICAL APPS KEEP THEIR PROMISES?

Raina Samuel, Iulian Neamtiu, Sydur Rahaman
New Jersey Institute of Technology

ABSTRACT

Medical mobile apps are already in wide use, and their use, as well as user base are projected to grow even further. However, it is unknown whether medical apps achieve their claimed behavior effectively and accurately. To determine potential gaps between app claims and app behavior, as well as between app claims and user expectations, we conducted a study on over 2,000 Android apps. We first developed an information retrieval approach that maps an app's description to medical (ICD) terms, hence delineate the app's medical scope and stated goals; our analysis has revealed that weight management, heart rate measurement, blood sugar measurement and hearing aids constitute the most common conditions apps claim to address. Next, based on app functionality, we categorize apps into (a) apps that measure or manage a physiological parameter, (b) apps that claim to treat conditions, and (c) apps for self- assessment. Within these three categories, we establish fine- grained subcategories and for each subcategory we compare apps' claimed behavior with realizable behavior. We found that app widely overstate their behavior and functionality. We also found that apps employ disclaimers and misleading terms to lure users into installing/using the app yet avoid responsibility. Finally, based on our uncovered app behavior and claims, we outline actionable findings w.r.t app claims and actual vs. stated function, meant to make users safer and apps more forthright.

KEYWORDS

Medical apps, Android, Digital health, Mobile computing

1. INTRODUCTION

Medical mobile apps are integral to daily life, offering diverse functions from connecting to medical devices to tracking physiological parameters to condition assessment or diagnosis.
Due to their convenience and ubiquity, users trust medical apps and generally assume that apps are validated and accurate. However, there is no direct evidence on whether a medical app is performing its claimed functions. For instance, in a study regarding blood pressure monitoring apps, users liked the perceived accuracy; however, the app under-reported users' actual systolic pressure and provided inaccurate results which gave users a false sense of security (Plante, 2018).

We conducted a study on more than 2,000 Android apps collected from Google Play to understand (1) the medical conditions targeted by medical apps, and (2) the claims app make, e.g., regarding diagnosis or cures; hence consequently reveal and categorize lapses between app claims and actual functionality.

To define app behavior and nature, we map app metadata terms onto ICD-11 (International Classification of Diseases) codes. Using ranked retrieval text analysis, we were able to accurately shed light on common conditions applications may claim to treat or manage (Section 2). For apps that perform measurement and tracking, we found that most ICD codes were related to physiological management, such as weight loss or heart rate measurement. For apps that address conditions, we found that the most common conditions include elevated blood glucose level (MA18.0) and speech therapy (QB95.5); we present the findings in Section 3.

Next, we focus on exposing questionable claims found in app descriptions.

We classified apps into three main categories of claimed behavior: physiological (Section 4), treatment (Section 5), and self-assessment (Section 6). Focusing on app descriptions allows us to observe better what may possibly convince users into installing certain apps. We establish keywords and frequencies to categorize suspicious behaviors accordingly. Within each category, we investigate app claims and compare these claims with what is realizable with an app running on a smartphone; we found a wide gap between claims and attainable functionality.

Overall, we make the following contributions:

- A classification of app behavior based on medical conditions established by international standards (ICD-11).
- A classification of possible misleading claims found in Medical apps.
- A discussion of app disclaimers and misleading description terms.

Prior work Several studies investigated the accuracy and overall usability of mobile health apps. Coppetti et al. studied the accuracy of smartphone heart rate measurement apps and revealed that there were substantial performance differences between heart rate apps and clinical monitoring -- as much as 20 beats per minute (Coppetti, 2017). While their study focused on only 4 apps, it is safe to assume that this issue persists with other apps of this nature.

The importance of responsible app marketplace safeguards regarding health apps is discussed by Wykes et al. (Wykes, 2019); their work expressed concerns with the "overselling of health apps" and suggested a set of four principles that app marketplaces could use to guide the user to more sensible choices. Their study was conducted on four apps, rather than a larger dataset. Wisniewski et al.'s study on top-rated health apps confirmed our findings that most medical apps "continue to have no scientific evidence to support their use" (Wisniewski,2019); their study is based on manual analysis of 120 apps.

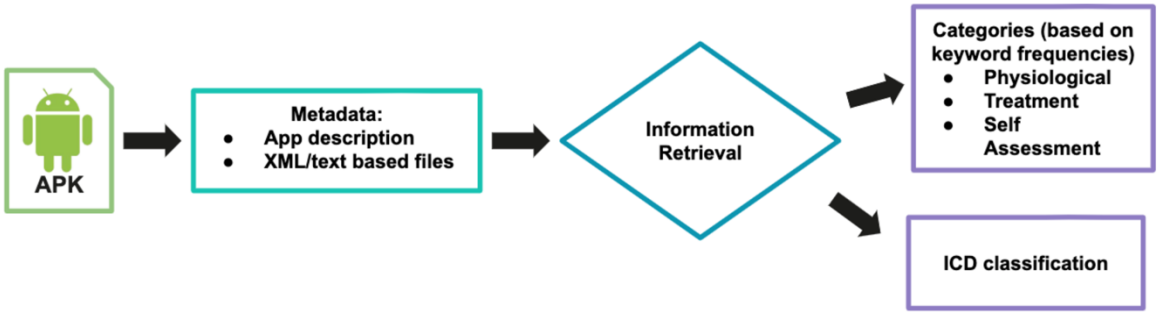
Other studies show that there is little evidence to whether health apps work, finding that only a small fraction of apps is tested (Byambasuren, 2018), leading to suggestions of "prescribed health apps", meaning having health apps vetted by medical professionals as a prescription rather than being able to be freely installed.

The reliability and safety of health apps is discussed by Akbar et al. with most concerns stemming from the quality of content presented in apps, such as presenting incorrect and incomplete information (Akbar, 2020).

Regarding weight loss apps, Zaidan et al. addressed the usability features of these apps by applying an evaluation framework (Zaidan, 2016). Their framework revealed that app marketplace search engines had biases towards certain titles and keywords that did not reflect the full functionality of the app and that the most popular apps are not necessarily the most effective.

Brown et al.'s review of 76 pregnancy apps regarding nutrition determined that such apps should not be considered as an appropriate resource for pregnant women due to unsound nutritional advice and overall unreliability (Brown, 2019).

Figure 1. Overview of methodology.



2. METHODOLOGY

We begin by describing our overall methodology illustrated in Figure 1. We acquired app descriptions and apps (APK files) from Google Play which correlated to the Medical and Health & Fitness categories. This resulted in a total of 2,339 apps that had installs over 1000 and were in the English language. From these apps, we used their descriptions and relevant text-based app metadata to determine ICD codes and observe claimed app functionality and misleading claims. We map the medical conditions apps may claim to treat (or monitor) onto an established ontology, ICD-11 codes. ICD -- the classification of diseases used by the World Health Organization -- provides an international standard for uniform naming of diseases and health conditions. Extracting ICD terms from app metadata not only enables us to identify possible conditions apps

may claim to treat or monitor, but also (1) can reveal lapses in app descriptions regarding functionality and (2) helps us understand further the general medical app landscape. App descriptions and text metadata were processed, removing stop-words and irrelevant terms. We will next discuss how we managed to extract ICD codes and categorize app functionalities via information retrieval.

ICD code mapping challenges

We begin by describing the process and challenges faced when extracting ICD codes from apps. We used 106 ground truth apps as a basis to determine the score threshold for matched terms.

The first challenge was determining relevant app metadata. Using irrelevant terms and certain stop words can result in inaccurate ICD mappings or even no matches. We compared extracted keywords between the app description and XML files in order to understand common medical app functionalities and which source would be the most effective in mapping with ICD codes. We found that XML files provided less relevant results despite having more medically related keywords, especially for apps used as reference material, intended for patients, or to interact with patient portals. This is because XML files contain more fragmentation and individual words rather than cohesive sentences to provide any meaningful input.

The second challenge involved finding the best method to map and extract terms from app descriptions. We began with an initial mapping with a TF-IDF (term frequency – inverse document frequency) analysis, which showed that most apps correlated to the ICD code MA13.1 (*Finding of alcohol in blood*). However, when we attempted a ranked retrieval text analysis, we found much more accurate ICD terms mapped to keywords.

Table 1. Keyword Discrepancies in ICD Codes

App Name	TF-IDF	Ranked Retrieval	ICD Code
com.ebsco.dha	'management',**'difficulty', 'disorder', 'condition'**	'health','refer','clinic','care'	QB10 (Medical services not available in home)
com.pocketprep.nptepta	'disease', **'specified', 'defect', 'vertical'**	'brain','test','therapy', 'nervous system'	MB72 (Results of function studies of the nervous system)
com.ninezest.stroke	'therapy', **'devices', 'malignant', 'miscellaneous'**	'stroke','therapy', 'speech',**'enhance'**	QB95.5(Speech therapy)
com.srems.protocol	**'harm','malignant', 'classified', 'miscellaneous'**	'region','clinic','treatment', 'cardiac arrest'	MC82.1 (Bradycardic cardiac arrest)

Using a naive approach leads to inaccuracy in text extraction, as we show such discrepancies in Table 1. Here we compare the keywords extracted from TF-IDF analysis versus those from ranked retrieval; the text in bold indicates inaccurate or irrelevant keywords that do not map to the accurate ICD code displayed. We see that ranked retrieval provided the most accurate results. Thus, we used ranked retrieval to obtain each app's ICD code. We will discuss our results in Section 3.

Categorizing claimed app functionalities

Next, we will discuss how we categorized app behavior. Here we focused solely on app descriptions, as they are the initial reasons why users download applications. We used 33 apps as ground truth, which we had manually determined as potentially misleading due to specific terms in their descriptions. We focused on generic terms such as "diagnosis", "entertainment purposes", "instant", and "camera" and applied a TF-IDF analysis on the full dataset, resulting in a subset of 1250 apps matching these criteria.

Once the subset was established, we then manually reviewed common patterns based on general functionality to create a categorization. We developed another set of keywords to categorize the 1250 apps into three categories using TF-IDF analysis. Finally, to better refine our categories and the broad functions we found, we further characterize them into more specific subcategories. In doing so, we reveal possible lapses in claimed behavior and their legitimacy, especially in popular apps. We further discuss our findings in Section 4.

3. MEDICAL CONDITIONS

We will now present our findings. First, we will discuss the results of the ICD code analysis and the top codes found. Then we will describe the claimed app behaviors found in app descriptions.

3.1 Top ICD Conditions

ICD codes extracted from app descriptions help us ascertain whether descriptions accurately describe/explain app functionality. Table 2 displays the top ICD codes along with an explanation of how it is used and its categorization.

We found that most of the ICD codes relate to weight loss apps due to the frequency of the term: MG43.5 (*Excessive weight loss*) as we see in Table 2. We also see many ICD codes related to apps which connect to external medical devices, especially with pacemakers (QBB30.3: (*Adjustment or management of vascular access device*) and hearing aids (QB31.4: *Fitting or adjustment of hearing aids*). Among the top ICD codes, we find very few apps for professional use relate to any, if at all. This is because many app descriptions related to professionals or clinicians are either very vague or too complex to map correctly to a single specific ICD code. Nevertheless, we were able to accurately map ICD codes to medical conditions in apps that claimed to treat said diseases.

Table 2. Top 10 ICD Codes based on app descriptions

ICD Code	ICD Title	#Apps	Use
MG43.5	Excessive weight loss	511	Weight Control
MC82.1	Bradycardic cardiac arrest	255	Heart Rate Measurement
QB30.3	Adjustment or management of vascular access device	242	Pacemaker Management
QB31.4	Fitting or adjustment of hearing aid	232	Hearing Aid
MA18.0	Elevated blood glucose level	135	Diabetes Management
M54.5	Low back pain, unspecified	120	Pain Management
QA41	Pregnant State	104	Pregnancy Tracking
CA23	Asthma	84	Asthma Management
QB95.5	Speech Therapy	75	Speech Aphasia Treatment
H93.1	Tinnitus	70	Hearing Aid

3.2 Claimed App Behavior

We characterized app functionalities into three main categories: Physiological, Treatment, and Self-Assessment. Apps in these categories are examples of behavior that may potentially require regulations or further scrutiny and exemplify the need to categorize claimed app functionality. Apps should be clearer about their true functionalities in their descriptions while being explicit in their disclaimers; many times, disclaimers are hidden in the text or towards the end of the Google Play description; when the description is lengthy, users may end up ignoring or missing the caveat completely.

We will now describe each category and subcategory found in Table 3 starting with Physiological, Treatment, and Self-Assessment.

Table 3. Categories of app functionalities

Category	#Apps	%
Physiological	430	34
Heart Rate Measurement	115	9
Optometry	98	8
Blood Sugar Measurement	87	7
Hearing Test	42	3
Skin Cancer	41	3
Body Temperature Measurement	31	2
Weight Loss	16	1
Treatment	320	26
Natural Home Remedy	200	16
Hypnotherapy/Brain Wave Therapy	71	6
Pain Relief	49	4
Self-Assessment	500	40
Mental Health	309	25
Symptom Tracking	106	8
Pregnancy Quizzes	85	7

4. PHYSIOLOGICAL

Apps in this category claim to be able to measure certain physiological parameters such as heart rate or blood pressure, using the camera and other smartphone sensors. Concerningly, these apps claim to provide some form of diagnosis based on the measurement; furthermore, the apps claim that their measurements are accurate. We have found 430 such apps, categorized as follows.

4.1 Heart Rate

Heart rate-measuring apps use the smartphone camera's flash feature to measure a person's pulse. Measuring heart rates via a smartphone camera is not inherently inaccurate or deceptive, though a study has found differences between results obtained with apps versus results gather via clinical monitoring (Coppetti, 2017). However, users should not solely rely on such apps for diagnosis or treatment. For example, app **Cardiac diagnosis (arrhythmia),** with over 1,000,000 installations, states no disclaimers or recommendations to seek a medical professional or use an actual heart monitor along with the app. The accuracy is generally unknown, especially how the app manages to detect such conditions. Unless these apps work in conjunction with an external medical device, such as a blood pressure meter or heart monitor, the accuracy of such apps should not be relied on for diagnosis. Moreover, we believe that (1) such apps should include a disclaimer or recommendation to consult a medical professional, and (2) the term ‘diagnosis' should be removed from apps' titles.

4.2 Optometry

Optometry apps claim to measure vision acuity by providing eye exams testing for astigmatism, near and far-sightedness or color blindness.

While these apps may provide a basic benchmark for vision, without a medical professional's diagnosis, the apps should not be used as a sole medical opinion. As a result, all apps with this functionality must include a recommendation to report their results to qualified ophthalmologists or optometrists before taking any sort of action.

4.3 Blood Sugar

Blood sugar apps claim to measure or track blood sugar. While many of these apps do have this behavior, as they work with a glucometer, many do not -- the apps simply serve as a journal.

Apps claiming to measure or track blood sugar without connecting to a glucometer or any sort of device can be misleading. Additionally, some apps whose name contains “Blood Sugar Test” have disclaimers stating the app cannot measure blood sugar but provides information on how to manage diabetes. Thus, these apps should modify their titles to better reflect app functionality, e.g., “Blood Sugar Tracking” or “Blood Sugar Log”.

4.4 Hearing Test

Hearing test apps are different from hearing aid apps, which tend to connect to an external hearing aid device, serving as a remote control. These apps claim to provide (1) tests regarding tinnitus and (2) therapies for hearing issues; nevertheless, users need to see an ENT or audiologist for a reliable and accurate diagnosis.

4.5 Skin Cancer

Skin cancer apps use the device's camera to take pictures of skin and then use an AI algorithm to provide a preliminary diagnosis regarding skin cancer. Apps claiming to detect skin cancer solely through a device's camera and without a blood test are deceptive and misleading. An example would be the app **Medgic** which uses AI to check for dermatological conditions or diseases by using the device's camera. While AI algorithms

have been able to detect conditions before, prognoses cannot be solely confirmed by a simple photo of one's skin -- other tests must be administered in order to make a conclusion. The app's description contains a disclaimer, albeit at the end, stating how the app is not a replacement for medical advice and that not all results are 100% guaranteed.

Another app, **Visus**, states that it is an experimental application that is publicly deployed and that its algorithm is "*30% more sensitive and precise than a conventional board-certified radiologist*".

4.6 Body Temperature

Body temperature apps claim to measure users' temperature, e.g., to detect a fever. However, this is ultimately misleading, as mobile devices do not have any means to measure temperature in their sensors. Instead, these apps serve as a mere journal to track user-inputted values for body temperature.

4.7 Weight Loss

Weight loss apps are numerous by nature, as seen in our ICD mapping. However, in this specific categorization we focused on apps that over-promise results within an arbitrary or unrealistic time frame or even "instant" results. We found that many apps do not urge the users to seek medical opinions prior to attempting weight loss. For example, the app **Lose Weight Fast at Home - Workouts for Women** with over 1,000,000 installs, claims that users following the app's regimen will lose weight in 30 days. However, there are no mentions in the app description of the influence other crucial factors such as diet, water intake, or genetic factors, have in weight loss.

5. TREATMENT

These apps claim to be able to cure diseases. We have found a total of 320 apps, falling into several subcategories.

5.1 Hypnotherapy/Brain Wave Therapy

These therapies are complementary forms of medicine (i.e., used to supplement traditional treatment methods). Apps in this category tend to not mention the importance of standard or clinically proven medical treatments to be used in conjunction with their suggested therapies. Hypnotherapy results are generally not clinically proven and may have adverse effects on users who are prone to epilepsy or other neurological conditions (Gruzelier, 2000). For example, the app **Atmosphere: Binaural Therapy Meditation** which has over 500,000 installations, states that it is able to "heal your DNA" with its guided breathing and meditation; nevertheless, the app description contains a disclaimer that the app is only for "entertainment purposes" and should not be a substitute for medical treatment.

5.2 Natural Remedy

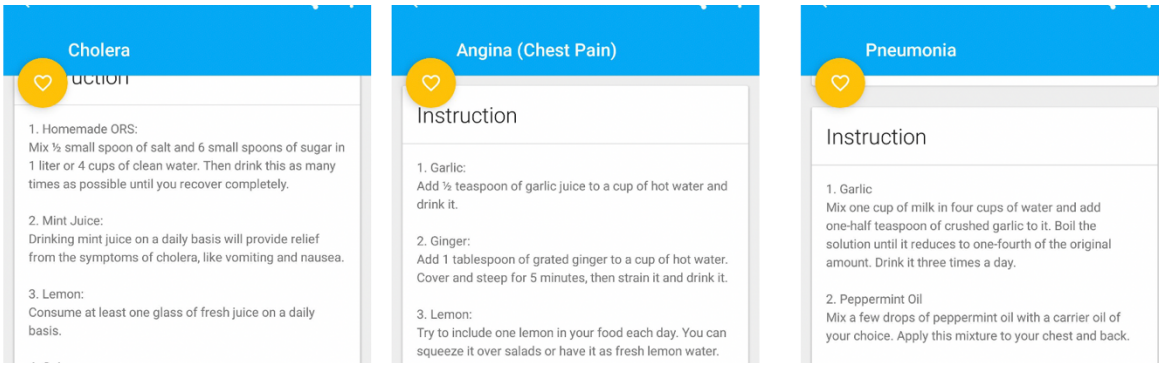
These apps provide references to natural remedies, e.g., certain herbs or foods, to manage and treat specific diseases, such as skin diseases or even cancer. They also claim to help users "self-cure" certain conditions. Apps that provide home remedies are not malicious or intentionally misleading but should never be a replacement for actual treatment prescribed by a medical professional. While there are natural remedies to basic non-life-threatening illnesses or wounds, an app is not an alternative to prescribed treatment from a medical provider (Desmet, 2004). For example, the app **Doctor at Home**, which has over 100,000 installations, claims it can provide treatment for "110 diseases" and "cure diseases at home". Examples of three conditions -- cholera, angina, pneumonia -- and app-prescribed "cures" are shown in Figure 2. Additionally, the app states that the user can be a home doctor, defined as "you are yourself a doctor". Herbal treatments and reference material cannot replace professional diagnosis or treatment. While the app has

useful tips for treating simple symptoms and issues, such as coughing and dandruff, it also has claims for treating more serious cases such as stomach ulcers and cholera.

5.3 Pain Relief

These apps rely on providing exercises and remedies to address various types of muscular pains or migraines. While such apps can offer a catalogue of exercises that can address certain types of pain, such apps should be used in conjunction with medical advice. Apps which work in tandem with qualified pain coaches can be a convenient way to help manage pain remotely. However, for many pain relief apps, pain is addressed through virtual exercises with claims that they are "proven to ease pain", such as in app **Lower Back Pain and Sciatica Relief Exercises**. Note that the issue is not whether exercises are effective or not; rather the issue is that app descriptions do not suggest seeking professional medical advice *prior* to app installation. Additionally, certain pain exercises, when performed incorrectly or without supervision, can lead to further damage and pain in many cases (Lubell,1989).

Figure 2. **Doctor at Home** claims to be able to 'cure' critical diseases naturally



6. SELF ASSESSMENT

We have found 500 apps which emphasize the use of assessments and self-help, categorized as follows.

6.1 Mental Health

Mental health apps rely on self-assessments without a professional entity providing feedback. Note that there is a lack of direct scientific evidence found in descriptions of apps that claim to help with mental health or behavioral patterns (Larsen, 2019). Many mental health apps do not provide confirmation or verification that the app is indeed vouched for by professionals. For example, the self-help app **MoodSpace** is focused on depression and mental well-being. While the description claims that the app is "a well-being app driven by research", there is no evidence of any research or authoritative proof accessible to users prior to installation. As with prior examples, the app's description contains a disclaimer and an emphasis that users should seek medical advice, but the disclaimer is found at the very end of the description, increasing the chance to be ignored by users.

6.2 Symptom Tracking

Symptom tracking apps are based on user input (rather than physiological measurements as prior discussed) to determine possible diagnoses. These apps are useful for a cursory understanding of certain symptoms but should not be used for a diagnosis. Many of these apps are extremely popular, such as **Ada-check your health**, with over 5,000,000 installations and classified as a Class I Medical Device, meaning it is considered as a device with low risk to the user in the European Union. While **Ada-check your health** is an example of a well-regulated medical app, there are many apps that claim to perform similar functions but

are not as well scrutinized or moderated by government or marketplace entities, such as the **Disease Detector** which claims to detect diseases in a few seconds.

6.3 Pregnancy Quizzes

These apps ask a series of questions and claim to determine whether the user shows early signs of pregnancy. While a collection of certain symptoms can help determine the likelihood of pregnancy, it can only be validated through an actual physical pregnancy test. As a result, the framing and naming of these apps are misleading. An example was app **Real Pregnancy Test & Quiz** -- removed from Google Play during this research -- which suggested that it was "*an easy quiz for pregnancy. Just reply the quiz questions*".

CONCLUSION

Medical mobile apps are understandably convenient and appealing to users. However, app quality and app description quality remain sorely lacking. These lacunae are particularly concerning in this (medical) domain because app reliability can directly affect/impact user safety and well-being. Our approach and study found that the functionality landscape of medical apps is broad and varied; however, the functionalities claimed in app descriptions are not entirely reliable. Our findings show a need for better regulation and scrutiny of medical apps in-app marketplaces to better protect users and their health.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their feedback. This material is based upon work supported by the National Science Foundation under Grant No. CCF-2106710.

REFERENCES

Akbar, S. et al. (2020) "Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review", *J. Am. Med. Inform. Assoc.* Oxford University Press (OUP), 27(2), bll 330–340.

Brown, H., Bucher, T., Collins, C. and Rollo, M., 2019. A review of pregnancy apps freely available in the Google Play Store. *Health Promotion Journal of Australia*, 31(3), pp.340-342.

Byambasuren, O., Sanders, S., Beller, E. and Glasziou, P., 2018. Prescribable mHealth apps identified from an overview of systematic reviews. *npj Digital Medicine*, 1(1).

Coppetti, T., Brauchlin, A., Müggler, S., Attinger-Toller, A., Templin, C., Schönrath, F., Hellermann, J., Lüscher, T., Biaggi, P. and Wyss, C., 2017. Accuracy of smartphone apps for heart rate measurement. *European Journal of Preventive Cardiology*, 24(12), pp.1287-1293.

Gruzelier, J., 2000. Unwanted effects of hypnosis: a review of the evidence and its implications. *Contemporary Hypnosis*, 17(4), pp.163-193.

Larsen, M., Huckvale, K., Nicholas, J., Torous, J., Birrell, L., Li, E. and Reda, B., 2019. Using science to sell apps: Evaluation of mental health app store quality claims. *npj Digital Medicine*, 2(1).

Lubell, A., 1989. Potentially Dangerous Exercises: Are They Harmful to All?. *The Physician and Sportsmedicine*, 17(1), pp.187-192.

Plante, T., O'Kelly, A., Urrea, B., MacFarlane, Z., Blumenthal, R., Charleston, J., Miller, E., Appel, L. and Martin, S., 2018. User experience of instant blood pressure: exploring reasons for the popularity of an inaccurate mobile health app. *npj Digital Medicine*, 1(1).

Wisniewski, H., Liu, G., Henson, P., Vaidyam, A., Hajratalli, N., Onnela, J. and Torous, J., 2019. Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *Evidence Based Mental Health*, 22(1), pp.4-9.

Wykes, T. and Schueller, S., 2019. Why Reviewing Apps Is Not Enough: Transparency for Trust (T4T) Principles of Responsible Health App Marketplaces. *Journal of Medical Internet Research*, 21(5), p.e12390.

"tfidf :: A Single-Page Tutorial - Information Retrieval and Text Mining" (no date). Available at: http://www.tfidf.com/.

Zaidan, S. and Roehrer, E., 2016. Popular Mobile Phone Apps for Diet and Weight Loss: A Content Analysis. *JMIR mHealth and uHealth*, 4(3), p.e80.