IEEE *Access*

A scoping review of the use of log data for evaluating mobile Apps: Exploring implications for mHealth Apps

A. van Schalkwyk[1], Prof. SS. Grobbelaar[2], Dr. E. Vermeulen[3], and Prof. M. Herselman[4]

[1]Department of Industrial Engineering, Stellenbosch University, Stellenbosch, 7600, South Africa
[2]Department of Industrial Engineering AND DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University, Stellenbosch, 7600, South Africa
[3]Department of Industrial Engineering, Stellenbosch University, Stellenbosch, 7600, South Africa
[4]Department of Industrial Engineering AND Next Generation Enterprises and Institutions, CSIR, Pretoria, 0002, South Africa

Corresponding author: S Grobbelaar (e-mail: ssgrobbelaar@sun.ac.za).

ABSTRACT There is a growing trend in the potential benefits and application of log data to evaluate mHealth Apps. However, the process by which log data is used to derive insights remains unstructured, resulting in the underutilization of mHealth data. We aimed to explore extant literature and guidance through a scoping review on how log data analysis can be used to generate valuable insights in support of the evaluation of mobile Apps. The scoping review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines for a scoping review. The Scopus database and grey literature (through a Google search) delivered 105 articles, and we applied inclusion and exclusion criteria to retain 33 articles in the sample for analysis and synthesis. This scoping review sought to identify how log data are used for mobile App evaluations. By highlighting the existing trends found in literature, identifying the similarities and differences between mHealth and General App analyses, and categorizing the indicators, insights, and improvements, this study contributes to the existing knowledge base of mHealth evaluations and future standardizations. The concepts and categories identified by this review are combined to form a proposed conceptual framework that will be refined and incorporated into future research toward addressing the gap identified in the current literature.

INDEX TERMS Evaluation, frameworks, mhealth, mobile Apps, scoping review.

I. INTRODUCTION

A. BACKGROUND

Digital technologies have permeated almost all domains of society, including health. Mobile Apps, as one such technology, support an array of everyday life activities (evident for instance in the broad range of App categories found in the Google Play Store). These Apps are developed in fast-paced agile environments, with numerous updates required while continued use of the App. When it comes to health-related Apps (mHealth) as healthcare interventions, there seems to be an incongruency between the fast-paced agile development of Apps and the more tedious traditional evaluation processes to establish healthcare interventions' efficacy and safety. The analysis of log data may alleviate this asynchrony. Forming part of the work towards a more comprehensive framework to incorporate log analysis into

Monitoring and Evaluation (M&E) for mHealth Apps [1] this scoping review also explores similarities and/or differences in how log data analyses are applied to General and mHealth Apps.

mHealth Apps propose several potential benefits that could aid effective assistance or improvements for healthcare delivery [2]. Extensive investment of resources is involved with developing and implementing mHealth Apps. However, very few mHealth initiatives proceed past the pilot phase [3], resulting in a loss of investment costs and potential benefits.

Additionally, with the vast number of mHealth Apps available on App Stores, it is difficult to determine which Apps are trustworthy and viable to implement for pilot or large-scale mHealth initiatives. This shows the need for evidence-based results and research [4]. The integration of M&E practices in the development and operation of mHealth

Apps would provide the needed inputs or 'evidence' to support the mHealth initiatives for sustained usage.

Traditional evaluation methods (such as Randomized Controlled Trials) are not suited to the iterative and fast-paced (i.e., agile) environment associated with (mHealth) App developments [5]. This warrants the need for improved and more suitable methods of evaluation for mHealth Apps. Although a range of frameworks have been developed or proposed for evaluating mHealth Apps [6–8], the evaluation frameworks' lack of consistency and comprehensiveness calls for standardization and improvements.

Reviews of the existing frameworks [6–8] highlight some of the challenges, such as the lack of comparators [8], concerns about adequately predicting engagement [7], and problems with vague or subjective assessment criteria [6]. To address such challenges, there is a call for more complete data and process transparency [7], with strategies such as app metrics or benchmark criteria to obtain accurate responses. Consequently, log data analytics would be a valuable contribution to addressing these challenges if utilized adequately.

Log data can be defined as "anonymous records of real-time action performed by each user" [9]. Log analysis (i.e., using log data to generate insights) could provide valuable inputs to support the functionality and usability aspects of mHealth evaluations. Log analyses provide the opportunity for real-time and objective information (or improvement points) about the technology and the process (user-technology interactions), making it suitable for formative evaluations. Log analysis could also explain the technology's uptake (i.e., the implementation and usage) and outcomes, which could assist the summative evaluations [9].

The authors acknowledge that log data analysis would mainly contribute toward one aspect of engagement: the micro engagement – which refers to the actual usage of the app [10] or the 'actual usage' aspects of adherence [11]. Micro engagement data do not necessarily reflect macro engagement – which refers to actual behaviour change of the users as a result of the app's usage [10]. Yet, these are still fundamental and valuable insights that would not have been possible to derive previously. Insights from micro engagement data (i.e., log data) could be used to define effective engagement or inform future qualitative studies (as part of mixed methods approaches) [11]. This paper proposes that more evaluations would incorporate app metrics if the log data concepts and applications were appropriately structured. This, in return, increases the possibility of continuous (or real-time) evaluations and the comparability of the real-world usage of interventions. Towards the structured or standardized application of log analysis, it is prudent to investigate its application in the current research domain.

The overview of the extant research will identify what are the key concepts in the field and where the gaps in the literature lie for potential improvement projects. A scoping review is an appropriate method in this regard. Scoping reviews are typically conducted when the researcher aims to: examine the methods through which research is performed on a specific topic; identify available evidence in a specified field; clarify key concepts; or identify and analyze knowledge gaps [12].

Previous studies have investigated and reported on the structured process required for implementing log analysis [9,13,14], the value that log analyses provided in the context of electronic health evaluations [11], and the consolidation of analytic indicators of engagement based on health Apps for chronic conditions [10]. The previous studies' approaches incorporate a realistic evaluation perspective meaning that the log data should be analyzed in context to identify the mechanisms of actions evident from the use of the technology which was applied towards achieving a specific outcome pattern. The context can be defined as "any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves" [15].

Realistic evaluation, as proposed by Pawson & Tilley [6], moves beyond the experimental evaluations of asking "what works?" or "does it work?". Instead, realistic evaluation focuses on the context in which an intervention takes place, considering the mechanisms through which change is affected to drive the achievement of specific outcomes (referred to as Context Mechanism Outcome Configurations (CMOC)) [16,17]. This review article follows this same evaluation perspective.

As evident from the process mining framework developed in the context of mobile commerce [13], log data concepts applied outside of the mHealth environment could also apply to the mHealth evaluations. This extended scope could contribute to the move beyond descriptive statistics often associated with analyzing and reporting mHealth log data [9]. In order to identify and incorporate valuable log data concepts or applications that could apply to mHealth (or improve the mHealth App analyses), this review has been widened to consider all mobile Apps as part of the eligibility criteria, with the findings applied to the specific environment of mHealth M&E.

B. OBJECTIVES

Motivated by the potential value that structured log analyses could provide for mHealth evaluations, this scoping review aims to identify and categorize the key concepts used in the existing knowledge base when analyzing the log data of mobile Apps. Unlike previous studies that focused on only categorizing the indicators (of engagement) [10], this study proposes that the categories could be more applicable to practice if they were based on the process mining approach [13] and thereby categorizes the insights and improvements in addition to the log data indicators.

Indicators are defined as the measurable and objective information or entities stored as part of the log data. To this

aim, the structured approach of log analysis, as proposed by previous studies [9], [14], is incorporated in the data extraction of this review to identify key concepts used for each stage of the process (*e.g.*, the collection, analysis, and interpretation stages). The findings would thereby inform the development of a conceptual framework [1].

As part of a scoping review, it is important to understand the application field and highlight the existing literature gaps. This includes identifying the observed trends regarding the most published or highly cited authors, publications, and countries affiliated with the research area. It can be used to inform future studies or recommendations following this scoping review (*e.g.,* if a more detailed systematic review is required).

The critical appraisal of the selected literature (or risk of potential bias) is not considered mandatory for scoping reviews [18], and is not conducted as part of this scoping review. Still, the potential bias is acknowledged, and a structured approach for conducting and reporting the scoping review is implemented, as explained in more detail in the Methodology section. Lastly, this scoping review's broad research question is: "*How are log data used in the M&E process of mobile applications?".* This question is addressed by considering different aspects of M&E, such as the evaluation perspective, focus area, approach, and context, as explained by the Data charting subsection.

II. METHODOLOGY

A. REVIEW PROTOCOL
Following the five steps of the Scoping Review Methodological Framework developed by Arksey and O'Malley, the aim is achieved by: (1) Identifying the appropriate research question(s); (2) Identifying relevant studies that have applied mobile log analysis; (3) Selecting studies from relevant databases; (4) Charting the data according to predetermined codes; and (5) Collecting, summarizing and reporting the results relating to (a) what log data indicators are used, (b) what insights and potential improvement are generated, and (c) what process or framework are applied when analyzing the mobile App(s)'s log data. [19].

The findings are reported per the checklist and guidelines of the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) [20]. Per the PRISMA-ScR guidelines, the principal investigator develops a scoping review protocol. The draft protocol is reviewed by the research supervisors and updated as required. The protocol defines the specific eligibility criteria that are used for identifying and selecting relevant literature. Furthermore, the search terms, screening process, and codes for the data extraction are formulated and refined.

B. SEARCH STRATEGY AND DATABASE
A literature search of the Scopus Database was conducted on the 5th of March 2021 using the Publish or Perish software [21]. Due to the selected scope (i.e., including all mobile Apps (not limited to mHealth Apps)), Scopus, an Elsevier database, was identified as an appropriate database for sourcing the existing literature. Scopus includes cited references, more than just articles (e.g., books and conference abstracts are also included), and includes journal titles that go beyond the biomedical disciplines (e.g., includes Health, Social, and Physical Sciences) [22]. Scopus also allows researchers to use search queries to structure or standardize the search terms across various sources [22].

TABLE 1
SCOPE-RELEVANT SEARCH TERMS

Scope	Search Terms
Data Analytics	"log data" OR "log"
Specific log data (Device)	"mobile application" OR "mobile App" OR "mobile device" OR "mobile intervention"
M&E	"criteri*" OR "apprais*" OR "indicators" OR "metrics" OR analy* OR "evaluat*"
Only for Grey Literature search	"Case study" OR "policy"[1]

The search terms, as provided in Table 1, were used. This resulted in 79 records being identified. The citation years ranged from 2010-2021, with an average of 25 citations per year (as calculated by the Publish or Perish software [21]). The papers were arranged by h-index while the first author conducted the Level 1 (title and abstract) screening. Inclusion and exclusion criteria were applied for selecting relevant literature for the database, as shown in Table 2. The publication date is limited to articles published after 2008 (the date of launch of the first App store), and the language is limited to English only based on the language capabilities of the authors.

TABLE 2
INCLUSION AND EXCLUSION CRITERIA

|	Reason for inclusion		Reason for exclusion
IC1	The application or use of log data	EC1	No log data applied or used
IC2	Evaluation of a mobile App(s)	EC2	Log data not relating to mobile Apps
IC3	Use of objective or automatic log data.	EC3	Analysis of self-reported or subjective data only
IC4	Full or short paper (not abstract only)	EC4	No Access (to full or short paper)
		EC5	Duplicate record
		EC6	Published before 2008
		EC7	Non-English

Any uncertainties of the Level 1 screening resulted in the study being included for Level 2 (full article) screening. Upon investigation, the Journal of Medical Internet Research (JMIR) included many publications in scope. It was searched separately for literature that could have been missed through the Scopus search. The JMIR database search and the Google search for grey literature identified an additional 26 records. The Google search utilized the same eligibility criteria with

---

[1] The additional terms were used to limit the Google search. This search was conducted after the Scopus search – with the assumption that the policy

and case study results would have been included by the results obtained in Scopus.

IEEE *Access*
Multidisciplinary : Rapid Review : Open Access Journal

additional search terms ("Policy" or "Case study") included. Various combinations of search terms were used to ensure scope-relevant documents were identified, as highlighted in Table 1.

An additional 13 records were identified from the included literature's reference lists. These references underwent and passed the Level 1 screening, which resulted in 44 records that were included for the Level 2 (full article) screening. From the Level 2 screening conducted by the first author, 11 documents were excluded with the reasons being documented, as shown in Fig. 1. This resulted in a final 33 papers selected for the data analysis and synthesis.
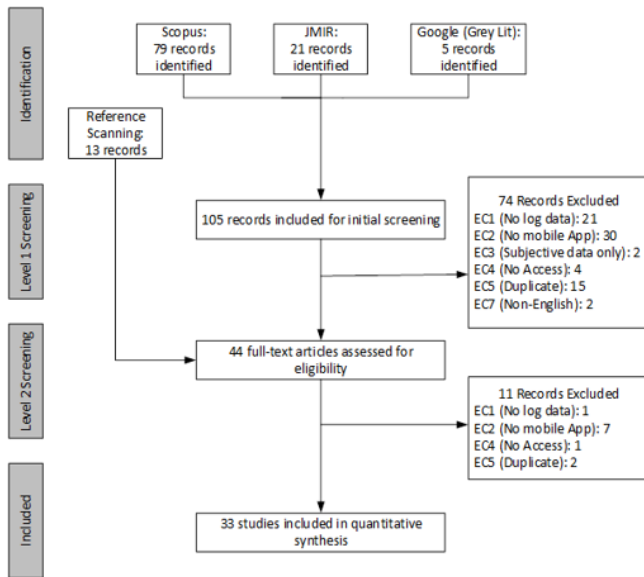
FIGURE 1. Literature selection strategy.

C. DATA CHARTING PROCESS

Following the study selection process, all selected literature is analyzed for data extraction using the qualitative data analysis software – Atlas.ti [23]. Atlast.ti [23] was used to identify and record the relevant codes evident in each study (according to the data extraction table, i.e., Table 3).

As shown in Table 3, the data extraction codes incorporated the context and methodology of both the research and the log analysis. This corresponds to the data extraction from a similar review [10], with additional codes focusing on the insights, techniques, and findings. The coding was expanded as new or valuable information was obtained, while the main focus was on the extracted data, as shown in Table 3.

The data extraction and analysis considered mHealth App (i.e., type of App) studies compared to the non-mHealth App studies (referred to as General Apps) to highlight any unique findings between the two categories. This also assisted in identifying potential areas or applications where General Apps' log practices could contribute to mHealth evaluations. The coded data are then analyzed using MS Excel to combine the codes highlighted in Atlas.ti into graphs and statistical values, as reported in the Results section. The findings are

divided into descriptive and conceptual analyses and reported accordingly using descriptive statistics (tables and graphs), narrative explanations, and quotes.

TABLE 3
DATA EXTRACTION CODES

Code	Description	Question
C1	Research Context	
1.1	Type of research	Does the literature have an empirical or theoretical focus?
1.2	Publication	Which journal or publication (*e.g.,* conference or website) is the article published in?
1.3	Geography	Where is the research conducted or what countries are the authors affiliated with?
1.4	Year	What year or timeframe is the study conducted?
1.5	Author	Who is the author(s) of the publication?
C2	Research Design	
2.1	Research Methodology	What research methodology is used to conduct the research?
2.2	Objectives	What are the stated objectives of the study?
2.3	Limitations	What limitations are highlighted?
C3	Log analysis Context	
3.1	Evaluation perspective	Is the evaluation or analysis conducted for accountability, for knowledge, or for development purposes?
3.2	Definition	How is log data defined?
3.3	Type of App	What type of App (*e.g.,* mHealth or General) is the log data gathered from?
3.4	Type of device	What type of device is the App operated on (e.g. phone, tablet, watch, etc.)?
3.5	First release	What year was the App first released?
C4	Log Analysis Methodology	
4.1	Process/framework for analysis	What process or framework is applied to conduct the log analysis?
4.2	Indicators	What indicators are gathered or monitored using log data?
4.3	Categories	In which categories can the indicator/insights be grouped in?
4.4	Insights	What insights are generated by using the log data (indicator(s))?
4.5	Tools/Platform	What tools or platforms are used to gather, clean, analyze, or store the log data?
4.6	Techniques	What techniques are used to gather/clean/analyze/store the log data?
4.7	Findings	What are the main findings identified from the research conducted?

III. REVIEW ANALYSIS AND RESULTS

A. DESCRIPTIVE DATA ANALYSIS

This section provides the findings from the descriptive data components as extracted from the C1 (Research Context) codes indicated in Table 3. It highlights popular publications,

geographical areas where the research is being conducted, trends in the year of publications, and the most cited authors.
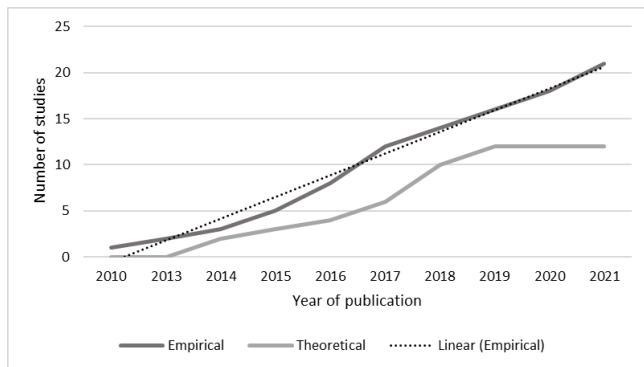

FIGURE 2. Cumulative publications per year based on type of research.

1) TRENDS OBSERVED PER YEAR OF PUBLICATION

As shown in Fig. 2, there has been an increasing trend of publications over the 2010 to 2021 period. The theoretical studies appear to be levelling out (following an s-curve), while the empirical studies show growth throughout. Due to the inclusion criteria of log analysis that had to be used (IC1), it is no surprise that more empirical studies were included in the review. However, it is interesting to note the cumulative number of theoretical studies included despite the inclusion criteria set.

2) SOURCES, AUTHORS, AND GEOGRAPHY

Sources were identified as journals (79%, 26/33), articles published in proceedings of conferences (15%, 5/33) and websites (6%, 2/33) (obtained from the grey literature sourcing). The selected literature came from 24 journals. The journals most commonly represented in our sample are Journal of Medical Internet Research mHealth and uHealth (18%, 6/33), followed by Institute of Electrical and Electronics Engineers (IEEE) Xplore (6%, 2/33), Journal of Biomedical Informatics (6%, 2/33), and the Journal of Interactive Marketing (6%, 2/33). The most represented geographical focus area (i.e., application area) is South Korea (21%, 7/33). This is followed by the United States (15%, 5/33) and China (15%, 5/33). It is noted that the majority of studies are conducted in developed countries. These findings indicate a gap in research for developing countries and the African region.

3) RESEARCH APPROACH AND METHODS

Research approaches can follow a deductive, inductive, or combined approach. The distribution of research approaches followed by the selected literature showed that 43% (14/33) followed an inductive approach, 30% (10/33) followed a deductive approach, and 27% (9/33) used both the inductive and deductive approaches.

The search terms and eligibility criteria excluded purely qualitative research methods as the log data had to be applied. The literature database was divided into Only log data analysis or Mixed-methods (if they combined the log data analysis with qualitative methods such as interviews or surveys). The

distribution of what methods are used for the analysis is then show annually (Fig. 3). Mixed methods studies seem to dominate since 2015 and have, in the past two years, been the only method applied. Mixed-methods were applied in 73% (24/33) of the papers (i.e., only log analysis was used in 27% (9/33) of the studies), which shows that most studies did incorporate qualitative methods when analyzing log data.
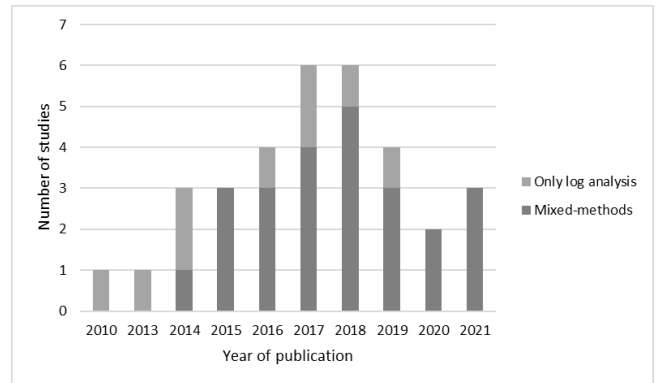

FIGURE 3. Number of studies that applied specific method of analysis per year of publication.

B. CONCEPTUAL DATA ANALYSIS

1) PROCESS, TOOLS, AND TECHNIQUES OF LOG ANALYSIS

The similarities and differences found in the current literature were explored to indicate to what extent consensus has been achieved regarding core concepts and approaches regarding how to conduct log analysis. Only 10/33 (30%) of the records articulated a conceptual framework, model, or process used to perform the log analysis. These are summarized as shown in Table 4.

TABLE 4
PROCESS OR FRAMEWORK FOR ANALYZING LOG DATA
SPECIFIED IN THE LITERATURE

Own framework/process specified	Reference
User Adoption Process Concept Model	[24]
Unique process specified	[25,26]
Process Mining framework	[13]
Application analysis platform	[27]
Own conceptual framework	[28]
3 Step usage path identification	[29]
Social mobile game concepts	[30]
Iterative process	[31]
Logging approach based on web usage mining	[32]

As shown, no study utilized the same framework or process. The remainder of the studies did not note a specific framework and used Exploratory analysis (52%, 17/33), Hypothesis testing (12%, 4/33) and Mathematical modelling (6%, 2/33). This correlates with the descriptive analysis findings that stipulated the popularity of the inductive approach. These findings confirm the lack of standardization in the research where log data is applied.

Different tools and techniques are available to conduct the selected process of log analysis. The selected literature mentioned 27 different tools used. Thirty-nine percent (13/33)

of the studies did not specify what tool was used. Among those specified: SPSS (5/33, 15%), R software (3/33, 9%) and Google Analytics (2/33, 6%) were the most popular tools. In this review, 21% (7/33) of the studies used only descriptive analysis (showing the mean, max, min, etc.) for analyzing the log data. The most popular technique used is data visualization (used by 39% (13/33) of the studies). More advanced statistical analysis (e.g., ANOVA, logistic regression, correlation tests) was used by 39% of the included studies (13/33) and both pattern analysis and Markov chain analysis by 6% (2/33). These findings indicate that more advanced data analytics (beyond descriptive analysis) are being applied to log data.

2) CONTEXT OF ANALYSIS

The context for log analysis takes into consideration the evaluation perspective, the focus of the analysis, the type of device or operating system, and the timespan of data being analyzed. These aspects, as found in the selected literature, are discussed in the following subsection. The studies are divided into mHealth App studies (52%, 17/33) or General App (i.e., all non-health Apps) studies (48%, 16/33) based on the type of App that was analyzed.

The purpose of analyzing the log data is divided into three possible groups according to the coded evaluation perspective (C3.1): Accountability, Development, or Knowledge. An accountability perspective focuses on the program's or intervention's results or efficacy; a development perspective uses the evaluative findings to strengthen the intervention; and a knowledge perspective aims to generate deeper understanding in a specific area, policy, or field [33].
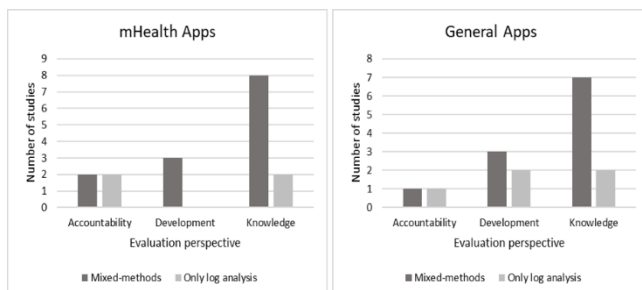
FIGURE 4. Number of studies that applied different method of analysis per evaluation perspective for mHealth and General Apps.

The differences in evaluation perspectives (cf., Fig. 4) between the types of Apps may be considered negligibly small. Therefore, inference on how perspectives are applied differently between the two App types is limited. This limitation is attributed to the small sample size. Yet, for the purpose of this article, each difference was noted and reflected on as discussed in Section IV.

The distribution observed (cf., Fig. 4) shows that the knowledge perspective is the most common evaluation perspective when analyzing mHealth and General Apps. The accountability perspective is more prevalent in mHealth studies, while the General Apps analyses have a higher occurrence of the development perspective. A mixed-methods

approach (i.e., using both quantitative and qualitative methods) is preferred for both mHealth and General App studies. Mixed-methods are more dominant for development and knowledge evolution perspectives, while there is no preference between mixed-methods or only log analysis for the accountability perspective.

Similar to the "engagement-related constructs" mentioned by [10], this review considered the specific focus area addressed by the log analysis. The studies included highlighted seven possible focus areas where log analysis of mobile Apps is applied. As diagrammed in Fig. 5, the four main focus areas are identified as Usability (40%), Engagement (15%), Effectiveness (15%), and Adherence (13%). Percentages for the Engagement and Effectiveness in our sample are similar, an observation which may be explained by how these concepts are related. Effectiveness refers to success or producing the desired result, which in the case of mHealth Apps is directly linked to how, how often and how long apps are used (usability and adherence) and what the usage ultimately results in (i.e., engagement).

This corresponds with the findings of a previous scoping review conducted on mHealth Apps [34]. The 'Other' focus areas (indicated in Fig. 5) include Simplicity of App (General App study) and Acceptability (mHealth App study) which had the lowest occurrences. Considering the focus areas split between the mHealth and General App studies, a similar distribution was observed for the Usability focus. However, adherence is more often associated with mHealth Apps, while Adoption was more popular with General Apps.
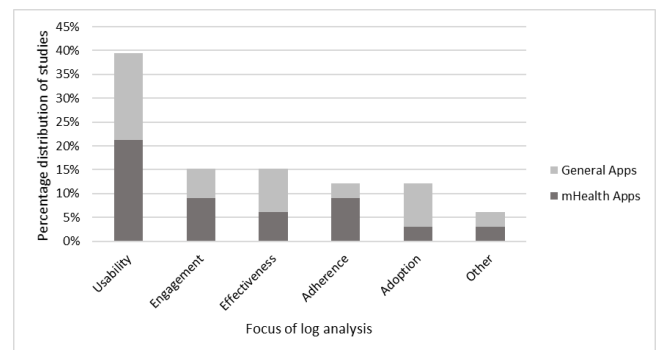
FIGURE 5. Percentage distribution showing the focus of the log analysis in the included studies.

The inclusion criteria (IC2) and the search terms meant that mobile Apps had to be part of the study. However, mobile Apps can operate on different devices, and devices can have different operating systems. This also forms part of the evaluation context, as different devices might have different indicators or concepts that form part of their log analysis. The literature analysis highlighted that 52% (17/33) of the studies mentioned: "mobile devices", including mobile phones, tablets, smartphones, and personal computers.

Only one study mentioned a logging device as part of its evaluation, and 15% (5/33) did not specify what device is used to run the App. Smartphones were mentioned by 8 of the 33

studies (24%), and tablets by 6% (2/33). The majority (55%, 18/33) of studies did not specify the operating system with Android operating system mentioned by 18% (6/33), iOS by 6% (2/33) and both operating systems by 21% (7/33). The document analysis highlighted that the device and operating system were mentioned to explain the App or the development description. It was used mainly for categorizing the user groups (e.g., view difference of results for Android users compared to iOS users).

The 33 studies highlighted an average of one year of log data used for the log analysis. The time from the first release to the time that the analysis took place is three years on average. Only two studies analyzed the App the same year as its first release. While the maximum years between the first release and the analysis was seven years. Thirty-six percent (12/33, 36%) of the studies did not specify when the App was first released. The maximum timespan of log data used was 5.5 years, while most studies used one month of data. Few studies (6/33, 18%) only analyzed log data collected during the study period (less than one month), and 2/33 (6%) of the studies did not specify the timespan of the log data analyzed or collected.

For effective evaluations, benchmarks or thresholds are required for comparability to determine whether the results are desirable or not and what could be improved. [36] state that there are no scales or standard measures for assessing the findings' relevance or comparing similar interventions and their results. This literature review confirmed this statement as no policies, standards, or set benchmarks were used to conduct or compare the log analysis findings.

The only benchmarks or thresholds that were mentioned or set (by 23/33, 67% of the studies) were with regards to classifying specific user groups (e.g. lost users, adopted users, engaged users, or active users) according to a specified period of use or non-use. Examples of the benchmarks set per reference can be viewed in the Appendix (Table A1). The user groups are classified by specifying the frequency of logins, intervals between usage, or duration of use during a set period. These benchmarks can differ according to the intervention type, goals set, evaluation purpose, or researcher's preference.

3) CONCEPTS AND INDICATORS
The initial data extraction chart (Table 3) aimed to extract the indicators associated with log data (C4.2). Upon further investigation, the indicators can be classified as 'collected indicators' and 'calculated indicators', in addition to the differentiation between mHealth and General Apps, and were extracted accordingly.

The collected indicators are used to derive the calculated indicators, which are analyzed to determine valuable insights (C4.4). These insights are used for future recommendations or improvements. Specific terms and concepts can also be grouped into categories or sub-concepts. Different terminology used for each concept is also highlighted, along with the various points of reference applied for analyzing and reporting the log analysis.
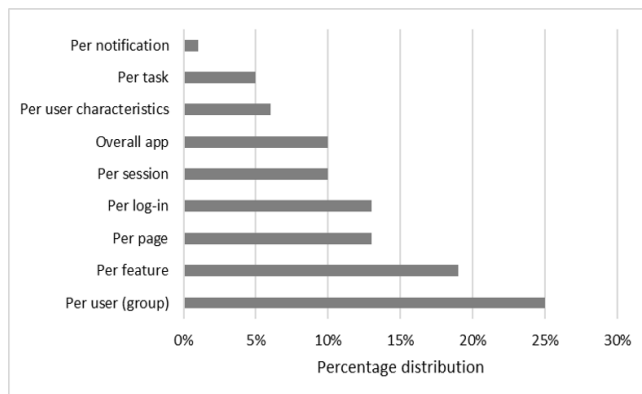

FIGURE 6. Percentage distribution of studies that mention the possible reference points used for log analysis.

Calculated indicators depend on the point of reference selected. The point of reference augments the analysis by aggregating calculated variables from the same perspective for better comparability of results. The point of reference can be divided into nine possible categories as found in the analyzed literature (cf., Fig. 6). Reporting the calculated indicators per 'user or user group' was the most used reference point. Again, these categories are not mutually exclusive. The results could include more than one reference; for example, the duration of use could be reported per feature and per user (user A spent two hours using feature Y). In addition, the timeframe selected also impacts the calculated indicators as the results can be reported since the launch of the App, annually, yearly, quarterly, monthly, weekly, daily, or hourly.

Identifying the concepts and categories highlights that similar concept are used by research studies but are often referred to using different terminology. The concepts used by this review identified the most popular or the most self-explanatory terms and then grouped the terminology according to the terms with similar meanings. The terminology found with their corresponding references are summarized in Table A2 (see the Appendix).
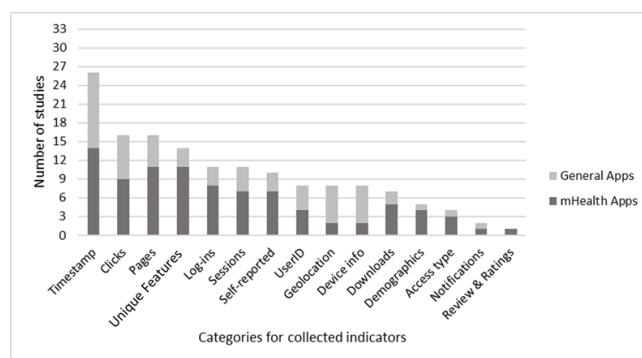

FIGURE 7. Number of studies incorporating collected indicators per App type.

Indicators collected as part of log data can be grouped into 15 different concepts, as shown in Fig. 7. Collected indicators

stipulate what is included in the log data and are collected based on the calculated indicators or insights that are desired. The description of each concept and the corresponding references that stipulated the collected indicator concept can be viewed in the Appendix (Table A3). The three main concepts most frequently included in the log data were: the timestamps (which consists of the dates and times) (26/33, 79%), each user event or click made (16/33, 48%), the specific pages (16/33, 48%), and the unique features of the App (14/33, 42%).

General Apps specified the collection of the geolocations and device information more often than mHealth Apps. In contrast, mHealth Apps were more likely to incorporate the specific pages, features, and self-collected measurements (e.g., blood pressure or goals set). A timestamp and the users' particular actions/events or 'clicks' would be required for any log data calculation (hence the popularity and equal distribution of the category). The low occurrence of some of the key categories (e.g., logins, userIDs, and sessions) is attributed to the fact that the collected indicators are not always explicitly stated.

Depending on what collected indicators were included in the log data, specific indicators could be calculated. The calculated indicators are provided as either the number, percentage, or statistical measures (e.g., mean, max, min, etc.) of a specific reference point during a set timeframe. Calculated indicators are not the final result but should be further analyzed to provide usable insights [5,35]. Five main conceptual categories are proposed to group the calculated indicators as found in literature: system (errors and reaction rate), notifications (notifications opened or received), usage patterns (location, retention, drop-outs, user properties, and sequence), time (intervals, peak periods, duration, and frequency) and features utilization patterns (popular or not used), as it works towards specific categories of insights (discussed in the following subsection).

The number of studies that included the specific calculated indicators is shown in Fig. 8, categorized into General and mHealth Apps. The corresponding references for each concept are provided in the Appendix (Table A4). Features utilization patterns (24/33, 73%), frequency (21/33, 64%) and duration (18/33, 55%) were the most common calculated indicators amongst all categories reported in the included literature. There is a low occurrence of retention rates (6/33, 18%) and drop-out points (7/33, 21%) presented in the included studies. The reaction rate of a notification sent (1/33, 3%) was unique to one study, while system errors, notifications opened/received, and locations were calculated by only two studies (2/33, 6%).

Although eight studies collected geolocations (Fig. 7), only two studies calculated users' specific locations. Most calculated indicators had relatively similar occurrences between the General and the mHealth Apps, except for User properties (that was majority mHealth Apps) and Drop-outs and Retention (that was majority General Apps). Compared to

the previous scoping review that found three different analytic indicators applied on average [10], this review identified similar findings as General Apps used on average three calculated indicators and mHealth Apps applied on average four.
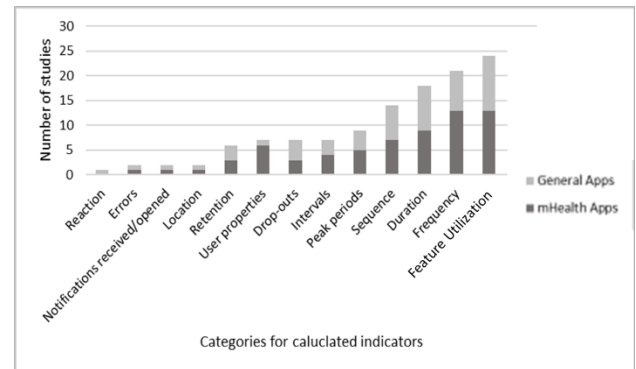

FIGURE 8. Number of studies that incorporated the calculated indicators.

[25] state that insights from log analysis can be considered in terms of user-level or feature-level insights. Based on the calculated indicators, the insights, and the recommendations made by the included studies, this scoping review proposes that, on a conceptual level, the insights be divided into 'user-level' and 'product-level insights'. User-level insights include existing, new, or potential user groups and user preferences. Product-level insights include insights about the technicalities of the App, the adoption, and the system influences. Insights are generated using the calculated indicator categories as mentioned previously. The link between what indicators lead to possible insights are diagrammed in Fig. 9.
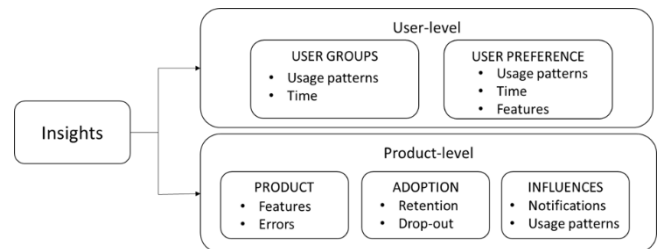

FIGURE 9. Division of possible insights generated from log data indicators.

User groups were identified according to the time-based [26,267,36], location-based [27,37], or device-based [27,36–38] frequency of usage. The time-based usage can be grouped into three main user groups according to the frequency of use: Active users (high-usage frequency), occasional users (medium or low usage frequency) and inactive users (non-usage), as summarized by Table 5. These user groups were evident in both mHealth and General App analyses.

TABLE 5
USER GROUPS WITH CORRESPONDING REFERENCES

User group	Different terminology	Reference
Active users	Active users	[30], [39], [40]
	High frequency	[28], [41], [42]
	Regular usage	[43], [44]
Occasional users	Occasional usage	[44]
	Medium frequency	[28], [39], [42]

	Low frequency	[28], [39]–[42]
	Used Sporadically	[43]
	Inactive usage	[30]
	Lost users	[24]
Inactive users	Not used	[39], [41], [43]
	Lack of use or discontinued use	[44]

Some researchers use the observed usage patterns to classify users into additional categories for future analysis, as summarized in Table 6. The users are classified according to benchmarks or threshold set for each group and App, which varies between research studies. Lastly, the user groups can be identified by categorizing or clustering users according to their: demographics, such as their age or gender [37,45]; their occupations or specialties [29,36,46]; or their device specifications such as the operating system [38], mobile platforms [37] or network [36]. User groups are identified and can be used to generate specific insights per user group. Based on the collected and calculated indicators, mHealth Apps were more prone to categorize users according to their properties or demographics, while General Apps were analyzed according to device information.

TABLE 6
ADDITIONAL USER GROUPS

Additional user groups for analysis	References
Engaged versus Disengaged	[31]
Adopters versus Non-adopters	[37]
Adheres versus Non-adheres	[40]
First time users versus Returning users	[13], [26], [30], [35]–[38], [47]

The user preferences, mentioned by 33% (11/33) of the studies, can highlight the discrepancies that might exist between intended use and actual usage, when the peak or popular usage times are, what features are preferred or used more often, and how the users respond to the intervention. The user preferences can be calculated per user group and contribute to product-level insights.

The product-level insights include insights about what features are used most often by whom, what features are not used, how the system errors could contribute to the feature not being used, how the usage (per feature) changes based on notifications sent or opened. It includes insights about how the retention rate changes per feature, per user group, based on the number of notifications sent, or how it compares to the desired retention rates. Key drop-out points can be identified, and insights can be drawn about how these could be avoided. Additional insights are also generated by comparing the feature usage, retention and drop-out rates, and notification reaction rates with the intended use or benchmarks set. This shows how the insights and indicators are not mutually exclusive but together form valuable insights.

The insights generated from the log indicators can be reported in terms of: the adherence to the intended usage [24,36,37]; the phases of the App's Adoption [24,36,37]; the Receptivity of following recommendations [46,48]; the Impact of a new feature or version [41,49]; the Effectiveness of the notifications [40,50]; the users' Engagement [51] or the

users' Satisfaction [29,35,42]. It depends on the focus of the research analysis and what indicators were initially collected.

The insights from the log analysis can be used to formulate valuable or actionable improvement suggestions [13,47]. The specific recommendations or improvement points will vary depending on the type of App, the focus of analysis, the indicators collected, and the insights generated. However, the potential improvement points can conceptually be grouped according to what insights contribute to the recommendations and whether it improves the App's usage or refines the log analysis. These proposed conceptual categories, their related research questions, and references are summarized in Table 7 and Table 8.

TABLE 7
CATEGORIZED POTENTIAL IMPROVEMENT POINTS FOR IMPROVING MOBILE APPS' USAGE

Aspect	Question	Possible improvement	Reference
Features	What can be done to improve the App or specific features?	Customie or personalie to suit user group or preferencezz	[4,12,24,35,51]
		Adding new features	[25,52,53]
		Removing non-used features	[52]
		Updating existing features (e.g., change position)	[30,53]
Adoption	How can retention be increased, or drop-outs be avoided? Should incentive or rewards be offered?	New or increased interventions, motivational strategies, or educational training	[12,29,38,46,51]
		If and when to send notifications/incentives	[39,49]
		Target specific user groups	[27,51]
Reduced Barriers	How can the app be made more compatible with its target user groups?	App compatibility with type of device (mobile, tablet, web), operating systems (iOS, Android, Windows)	REFERENCE TO BE ADDED
		Geolocation should inform language functionalities provided by App	

Aspect	Question	Possible improvement	Reference
Features	What can be done to improve the App or specific features?	Customize or personalize to suit user group or preference	[5,13,25,36,52]
		Adding new features	[26,53,54]
		Removing non-used features	[53]
		Updating existing features (e.g., change position)	[31,54]
Adoption	How can retention be increased, or drop-outs be avoided? Should incentive or rewards be offered?	New or increased interventions, motivational strategies, or educational training	[13,30,39,57,52]
		If and when to send notifications/incentives	[40,50]
		Target specific user groups	[28,52]
Reduced Barriers	How can the app be made more compatible	App compatibility with type of device (mobile, tablet, web), operating	[24], [26] –[28], [32], [35], [36], [47], [59]

IEEE *Access*

with its target user groups?	systems (iOS, Android, Windows) Geolocation should inform language functionalities provided by App	[25], [27], [32], [36], [37], [45]	

Table 7 shows that to improve the App's usage, either the Features or the Adoption can be improved, or the barriers of app usage can be reduced. Nine possible improvement points could be identified based on recommendations or applications in the current literature. The insights can also be used to refine the log analysis for future evaluations of the same or other Apps (cf., Table 8).

This is done by incorporating user group insights, additional qualitative research methods, or updating and documenting the benchmarks set for comparison. The analysis and improvement process occurs iteratively. Deviations (between the actual usage and the benchmarks set for the intended use) are reduced by either motivating the users or re-evaluating the benchmarks set [31,32] then observing if and how the results change.

Majority of customization improvement concepts were highlighted from mHealth App studies (4/5), along with the potential improvement of incorporating additional qualitative measurements (7/8). General App studies were the main contributors of persuasive triggers to increase adoption (4/5) and suggestion of updating benchmarks to suit actual usage/preferences (4/4). This correlates to the identified focus areas as General Apps focused more on Adoption and mHealth Apps focused more on Adherence.

TABLE 8
CATEGORIZED POTENTIAL IMPROVEMENT POINTS FOR
REFINING THE MOBILE APPS' LOG ANALYSES

Aspect	Question	Possible improvement	Reference
Benchmarks	How can the standards or benchmarks be more suited to actual use?	Change intended use to be more suited to user-level insights	[13,27,31, 32]
User groups	How can the user groups be used for additional insights?	Identify unique user groups from actual usage patterns	[24], [26], [51]
		Identify user groups based on unique properties (e.g., occupation, gender, age)	[29,36,45, 46]
Qualitative methods	How can insights from log analysis be used to add more value?	Conclusions or reasons for the observed usage can be identified by combining findings with qualitative methods (*e.g.,* setting up questionnaire about why specific features aren't used).	[25], [37], [39], [41], [47], [51], [55], [56]

4) FINDINGS AS REPORTED BY THE INCLUDED STUDIES

The findings as coded (C4.7) from the 33 included studies highlighted the value of log analysis. The two most important aspects of mHealth and General Apps' log analyses are the context and timeframe of the analyses. The context should always be considered when analyzing the log data [50] or providing app-specific recommendations [51]. This corresponds with previous studies investigating and affirming the importance and challenges of considering the context and context-awareness during health technology's development and use [15,57].

Within a specific context, the degree of satisfaction, the features used, or the particular usage behaviour are determined [44,53,56], which adds to the challenges of analyzing Apps within context as analysts need to make multiple decisions to decide and justify what is relevant and how the collected data will be applied [15]. The context, as defined during the Background section, is incorporated by the included studies in terms of: the target user groups (their lifestyle, demographics, or characteristics); the specific devices or platforms used [35,36,38,42,54,58]; and the previously discussed context of analysis (e.g., type of App, evaluation perspective, and focus area).

[25] highlight the usefulness of log analysis by explaining that the meaningful and timely insights meet the interventions' evaluative needs; however, the "data is significantly more useful when it is graphed over time". [56] state that the "results emphasized the importance of timing, tailoring, and ease of use". The findings highlighted those insights from log analysis would not have been evident from traditional evaluation methods [5] or subjective user opinions [55] and that the results could be used as valuable benchmarks for future evaluations [38].

The limitations identified from the included literature highlighted that the lack of structure (guidance) or benchmarks for developing insights from log analysis limits its use. Several studies reported the lack of or concern regarding the generalizability of their findings [26,27,35,36,42–45,53]. The remaining studies in our sample did not provide any structure to the process followed which, points to potential issues regarding the reproducibility of their findings [30,40,41,54,59].

Using only log analysis is often criticized about the accuracy thereof or the consideration of the 'dose-response aspect', as using more features or spending more time does not necessarily mean more engagement or better outcomes [58]. Some studies only mention the calculated indicators with no insights or potential improvement points generated beyond the descriptive statistics of usage [31,39,45,55,56]. Other studies also suggested that the limitations of log analysis can be addressed by incorporating qualitative methods with the findings generated from the log analysis [29,36,38,40,44,52].

IV. DISCUSSION OF RESULTS AND FUTURE RECOMMENDATIONS

This scoping review sought to identify how log data is being used for mobile App evaluations. By highlighting the existing trends found in literature, identifying the similarities and differences between mHealth and General App analyses, and categorizing the indicators, insights, and improvements, this study contributes to the existing knowledge base of mHealth evaluations and future standardizations.

A. TRENDS OF THE EXISTING KNOWLEDGE BASE

The results show an increasing trend in the publications within this scope. However, a clear gap is observed in the research conducted and published within developing countries. Developing countries, often associated with severe resource constraints, could benefit from implementing and sustaining mHealth initiatives [2]. Consequently, there exists an opportunity for log analytics research in developing countries' unique contexts.

Currently, there is a lack of standards for both the application and the reporting of mobile Apps' log analyses. The popularity of the inductive approach (43%, 14/33) and exploratory analysis (52%, 17/33) is attributed to the lack of theories, standards, and standardized frameworks or processes to follow when analyzing the log data. This was evident as only 30% (10/33) of the studies explicitly mentioned a framework or process used to analyze the log data, and none were used by more than one reference. Furthermore, different terminology was identified for similar concepts, and only 23/33 (67%) referred to set benchmarks. The field would thus benefit from a standardized framework to guide the collection, analysis, and interpretation of the log data towards standardization and comparability of the results.

Despite the short timeframe of log data collected and analyzed (one month), the log data seems to provide many potential insights. This allows for analyses to be done quicker than traditional methods (such as highly subjective user surveys) as prolonged data collection periods may be circumvented. The short timeframes used raise the question of how the collection, management, analysis, and insights of more extensive (i.e., collected over more extended periods) log data sets would change the findings of the studies in our sample. Log data can quickly form complex and large data sets (i.e., become Big Data). These datasets in itself could also be a valuable future research topic.

B. REFLECTING ON THE SIMILARITIES AND DIFFERENCES OF MHEALTH AND GENERAL APPS' LOG ANALYSES

This scoping review highlighted that regardless of the study analyzing mHealth Apps or General Apps, the mixed-method approach has been utilized over using only log analysis. Mixed methods are often used or proposed as benefits are drawn from insights generated using quantitative and qualitative methods, albeit it may be time-consuming. This review aims to contribute to the standardization of the log analysis (quantitative method) so that it can more easily be applied and

incorporated with the qualitative methods (for timely evaluations more suited to the App environment).

Planning for the analysis is required to determine what data needs to be captured to generate the desired insights, with particular consideration for the context of analysis. Although the knowledge perspective was the most popular evaluation perspective for all types of App evaluations (for mHealth and General), mHealth App analyses had higher occurrences of the accountability perspective than the General App analyses, which focused more on the development perspective. Similarly, the log analysis of all Apps focused on the Usability aspects, while Adherence was focused on more by mHealth Apps and Adoption by General Apps. Consequently, these evaluation perspectives and focus areas influenced the differences between the log data indicators and reported potential improvements.

Similar distributions of collected and calculated indicators were identified with a few exceptions, as discussed. mHealth Apps focused more on collecting the 'unique features', 'pages', and 'self-reported' indicators and consequently reported more User-level insights. These indicators contributed to the user groups formulated based on the user characteristics and demographics. The recommendations aimed toward 'increased training or motivations' and 'additional qualitative research' were reported more often by mHealth studies. These preferences are expected when focusing on adherence and establishing accountability. The purpose of mHealth Apps is often associated with behavioral change models that require specific outcomes to be linked to intervention and health impacts – hence the importance of following the intended usage (Adherence), proving the App is used as intended (accountability) and motivating/training the users (improvements) to achieve the intended usage or benchmarks.

In contrast to mHealth Apps, the General Apps reported most of the 'geolocations' and 'device information' collected indicators and a majority of the 'retention' and 'drop-out' calculated indicators. General Apps identified user groups based on device information and insights related to the Product-level categories. Thereby, 'customization' and 'changes to the benchmarks' were improvement categories associated more with the General Apps. Again, this links to the development perspective and Adoption focused analyses. General Apps are more profit-driven, thereby needing improved Adoption by monitoring and recommending ways to increase the number of users. General App developers stay focused on the device information to prevent Apps from becoming obsolete or irrelevant and ensure users are able to access and use the Apps on the intended devices (and operating systems, versions, etc.).
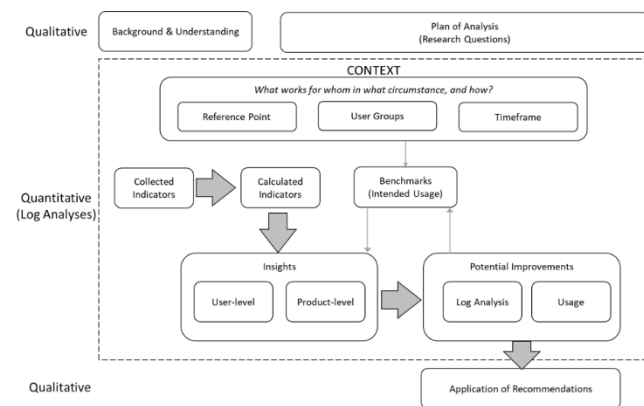
The similar division between mHealth App (52%, 17/33) and General App (48%, 16/33) studies and the similar distributions between the conceptual concepts identified show the benefit and popularity of log analyses for different contexts and Apps. It shows the multidisciplinary nature of the field of

study, where the various fields can learn from each other. For example, mHealth studies could incorporate development perspectives towards using the insights for App improvements. Additional indicators could be incorporated, such as geolocation and device information. Continuous and iterative monitoring and improvements to the technology's technical capabilities are essential considerations for any App environment (e.g., considering the App version, device capabilities, or difference in operating systems).

The mHealth studies could also consider which benchmarks or 'intended usage' are not crucial to the intervention impacts that could be adjusted to suit the user's preference – these customization recommendations relate to the emerging field of personalized medicine. There are challenges regarding the reliability and ethical concerns of adopting some of these practices. However, structured policies and standards could address these challenges and be valuable topics for future studies or projects. Similarly, General Apps could also learn from some of the indicators or insights of mHealth Apps.

Future recommendations could include more detailed systematic reviews on some of the key areas highlighted in this scoping review. Further developments based on the proposed concepts and categories could work towards a conceptual framework for structuring how log analyses are applied during mHealth or General App evaluations. Researchers or analysts could apply some of the recommended concepts and report on the feasibility of using cross-discipline indicators. Each App will have unique insights and recommendations, although these can also be grouped according to conceptual categories as presented in this review. By identifying and categorizing all possible concepts, the most suitable concepts to the App and its specific context can be selected, and benchmarks can be explicitly stated. This would result in practical improvement points and improve the comparability of the results. These terms and concepts are grouped, as shown in Fig. 10, to demonstrate how a conceptual framework can be developed and applied in future research studies.

Fig. 10 shows how the concepts identified within this review are interlinked in the overall mixed-methods approach. It demonstrates the process and considerations required for

structuring the log analyses and highlights the importance of gaining an appropriate background understanding. Lastly, the realistic evaluation principles are also incorporated, emphasizing the context considered throughout the entire process. This framework should be refined and tested in practical applications to determine its feasibility within the field of study.

V. LIMITATIONS

This literature review only included English documents with only one researcher selecting and extracting the data – this could contribute to a potential bias. The researcher aims to minimize the publication and literature bias by following a structured approach, documenting the entire process, and including more than one search database. However, the use of the JMIR database may have over-represented the number of mHealth Apps within the identified scope, skewing some of the results towards more focused mHealth insights instead of for all mobile Apps.

Valuable grey literature principles and practices may have been excluded based on the search terms and eligibility criteria focused on academic publications. Using Google as a search engine for grey literature also has some limitations. Google results vary depending on the location (country) in which they were searched and the previous search history [60]. Limitations and potential biases associated with the characteristics of scoping reviews are acknowledged. The small sample size is a limitation that could contribute to bias or inconclusive results.

Future studies such as systematic reviews with more than one reviewer, refined research questions, and the inclusion of quality assessments or critical appraisals of the literature are recommended to address these challenges. Lastly, the current literature trends favor time-based evaluations or insights, while the location-based and device-based considerations could also contribute to Apps' evaluation and/or engagement aspects– this should be noted and explicitly incorporated during future studies.

VI. CONCLUDING REMARKS

This review aimed to obtain an in-depth understanding of how log data analysis can generate valuable insights into mobile Apps. This aim was achieved by following the Scoping Review Methodological Framework [19] and documenting each step. Thirty-three documents were reviewed and analyzed by following the PRISMA-Scr guidelines [20].

The findings are reported according to the descriptive and conceptual analysis conducted. The descriptive analysis highlighted current trends and gaps in the existing literature, while the conceptual review provided an overview of key terms and concepts applied when analyzing Apps' log data. The review highlighted an overall lack of standardized terminology, processes, frameworks, and explicit benchmarks. Thereby, the need for a conceptual framework that is able to standardize the log analysis of mobile Apps is

IEEE *Access*

Multidisciplinary : Rapid Review : Open Access Journal

determined. Finally, the concepts and categories identified by this review are combined to form a proposed conceptual framework that will be refined, incorporated, and applied in future research toward addressing the gap identified in the current literature.

APPENDIX

TABLE A1
EXAMPLE OF BENCHMARKS SET WITH CORRESPONDING REFERENCES.

Category	Benchmark	Reference
Lost users	30 days not used	[24]
	7 days no log-ins	[25]
Adopted users	Downloading and logging in at least once	[25], [37]
	Used system at least once a month	[29]
	Minimum intervention dosage is one hour per week	[35]
	Intended usage is 3 hours per week	[40]
	App is accessed within 7 days	[51]
Frequent/Active users	More than 50 keystrokes per month for at least 6 months	[43]
	More than 5 user inputs	
	App opened more than 30 times a day	[50]
	Log-in more than 4 days per week	[41]
	Logged in more than 5 times	[49]
	Logged in one or more times per day	[28]
	Used App at least one time during the week	[39]
	Used App for more than 3 hours during the day	[42]
	App used at least once per month	[38]
	Recorded at least two entries – more engaged users self-reported higher number of conditions	[45]
	Several times each week	[30]
Inactive/Low-frequency users	One or no App openings per day	[50]
	Log-ins less than 4 days per week	[41]
	Used App for less than one hour per day	[42]
Non-users	No log-ins during trial	[41]
	Less than once a week	[28]
End of session	Longer than 30 min of inactivity	[5]
	Longer than 60 min were excluded; Logs out automatically after certain time of no activity	[29]
	Automated timeout of 10 minutes of inactivity	[35]
Norms/standards	Fitness norms; Biological age	[44]

TABLE A2
CONCEPTS AND TERMINOLOGY FOR LOG DATA WITH THE CORRESPONDING REFERENCES

Concept	Additional terms	Reference
Features	Activities	[13], [26]
	Events	[25], [26], [32], [38], [47]
	Category	[49]
	Intervention tools	[58]
Timestamp	Time that feature is click-on	[25]
	Time of occurrence	[13]
	Event time	[36]
Clicks	Selections	[55]
	Performed Clicks	[5]
	Action	[32], [35], [40], [44], [53]
	Taps	[32], [52]
Log-ins	Access	[49], [51]
	Visits	[13]

Geolocations	Location information	[27], [45]
	Internet protocol address trail	[25]
	Latitude and longitude	[50]
	Location check-in	[37]
Start time	Time of first check-in	[25]
	Start of use	[5]
	Time of access	[52]
End time	Exit time	[24]
	Exit of App	[5]
	Last log-in	[30]
	Final click	[35]
Sequence	Clustering of different sequences or trials of user actions	[5], [13], [41], [45]
	Navigation from one content area to another	[32], [38]
	Move to next screen, page or feature	[13], [28]
	Usage patterns	[36]
	Usage path of adjacent features selected (process discovery)	[29], [32], [35], [55]
Sequence	Order of actions	[40], [57]
	User trace (mix of activity patterns)	[26]
Duration	Time spent	[26], [28], [30], [35], [38], [40], [58], [59]
	Stay time	[13]
	Exposure	[35]
	Access period	[36]
	Session length	[38]
	Longevity	[45]
	How long	[58]
Drop-out rate	Loss rate	[24]
	Drop-out percentage	[41]
	Discontinued use	[38]
	Inactiveness	[30]
	Churn rate	[31]
	Ceasing to use	[58]
Retention rate	Stickiness	[37]
	Continued use	[24], [42]
	Rolling retention & return rate	[38]
	Percentage of end-users	[46]

TABLE A3:
REFERENCES CORRESPONDING TO COLLECTED INDICATOR CONCEPTS.

Collected indicators	Description of concept	References
Clicks	The user taps or actions being captured	[5], [24], [26], [28], [31], [32], [35], [38], [40], [43], [44], [49], [52], [53] - [56]
Log-ins	The specific event of entering or starting the App	[31,35–37,40,41,44,49, 51,58,61]
Self-reported measurements	Information entered by the user (*e.g.,* blood pressure, goals, progress)	[25–27, 30, 35, 37,38,40,43,45]
User ID	Anonymous unique identifier assigned to the specific user	[30], [32], [33], [40], [42], [45], [50], [56]
Timestamp	The date and time of the specific action or click being recorded	[5], [13], [25], [27], [28], [32], [35], [36], [40], [46], [47], [53], [56]
Session identifier	Indicates unique sessions of use – includes the start and end time	[5], [26], [28], [29], [32], [38], [40]
	Start time	[5], [24], [25], [30], [40], [44], [50], [52], [56]

	End time	[5], [24], [30], [35], [56]
Geolocation	The coordinates (latitude and longitude), location information, or IP address/trail of use	[25], [27], [32], [36], [37], [45], [50], [59]
Unique Features./Pages used	The unique functions or activities provided by the App that determines the layout	[25], [26], [29], [32], [35], [36], [38], [40] -[44], [47], [49], [57]
Access type	Means of entering the App or starting use (first page view)	[36], [46], [50]
Number of downloads	Number of times the App has been downloaded or installed	[24], [36]–[38], [40], [41], [49]
Reviews and ratings	The subjective reviews and ratings collected from the users	[38]
Device information	Includes the carrier type, target device, type of cellular network, operating system, App version	[24], [26]–[28], [32], [35], [47], [49], [59]
User characteristics	Includes date of birth, gender, user occupation or role, age, etc.	[26], [35], [36], [45], [46]

TABLE A4: REFERENCES CORRESPONDING TO CALCULATED INDICATOR CATEGORIES.

Calculated indicators categories	References
NOTIFICATIONS	
Notifications received	[50], [58]
Notifications opened	[50], [58]
SYSTEM - omission errors	[52], [53]
FEATURES	
Feature utilization	[5,13,26–31,35, 38–40, 42–46, 49–53,58,59]
TIME	
Frequency	[5,13,27–29,31,35–37, 40–47,49,52,56,58]
Interval	[26,27,36–38,40,42]
Duration	[13,26–28, 30, 32, 35, 36, 38, 40, 41, 44, 45, 51, 52, 56, 58, 59]
Time of day	[5,24,25,30,38,52]
Reaction	[32]
USAGE PATTERNS	
Sequence	[5,13,26,28,29,32,35,36, 38,40,41,45,55,58]
Retention	[24], [27], [37], [38], [40], [42]
Drop-outs	[24], [30], [31], [37], [38], [41], [58]
User Properties	[27], [29], [30], [36], [40], [44], [45]
Location	[50]

ACKNOWLEDGEMENTS

ABBREVIATIONS

Context Mechanism Outcome Configurations (CMOC)
Journal of Medical Internet Research (JMIR)
Monitoring and Evaluation (M&E)
Mobile health application(s) (mHealth App(s))
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR)

REFERENCES

[1] A. van Schalkwyk, S. Grobbelaar, E. Vermeulen, and M. Herselman, "Contributing to monitoring and evaluation practices by conceptualizing a Valuable Insights From Log Analysis (VILA) Framework," 2022.

[2] A. Botha and V. Booi, "MHealth implementation in South Africa," 2016. doi: 10.1109/ISTAFRICA.2016.7530667.

[3] R. Pankomera and D. van Greunen, "A model for implementing sustainable mHealth applications in a resource-constrained setting: A case of Malawi," Electronic Journal of Information Systems in Developing Countries, 2018, doi: 10.1002/isd2.12019.

[4] P. Henson, G. David, K. Albright, and J. Torous, "Deriving a practical framework for the evaluation of health apps," Lancet Digit Health, 2019, doi: 10.1016/S2589-7500(19)30013-5.

[5] J. Stragier et al., "Data mining in the development of mobile health apps: Assessing in-app navigation through markov chain analysis," J Med Internet Res, 2019, doi: 10.2196/11934.

[6] M. Hensher et al., "Scoping review: Development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers," Journal of the American Medical Informatics Association, vol. 28, no. 6, 2021, doi: 10.1093/jamia/ocab041.

[7] S. Lagan, L. Sandler, and J. Torous, "Evaluating evaluation frameworks: A scoping review of frameworks for assessing health apps," BMJ Open, vol. 11, no. 3, 2021, doi: 10.1136/bmjopen-2020-047001.

[8] M. R. Moshi, R. Tooher, and T. Merlin, "Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: A systematic review," Int J Technol Assess Health Care, 2018, doi: 10.1017/S026646231800051X.

[9] F. Sieverink, S. Kelders, M. Poel, and L. van Gemert-Pijnen, "Opening the Black Box of Electronic Health: Collecting, Analyzing, and Interpreting Log Data," JMIR Res Protoc, vol. 6, no. 8, 2017, doi: 10.2196/resprot.6452.

[10] Q. Pham et al., "A library of analytic indicators to evaluate effective engagement with consumer mHealth apps for chronic conditions: Scoping review," JMIR Mhealth Uhealth, 2019, doi: 10.2196/11941.

[11] F. Sieverink, S. M. Kelders, L. M. A. Braakman-Jansen, and J. E. W. C. van Gemert-Pijnen, "The added value of log file analyses of the use of a personal health record for patients with type 2 diabetes mellitus: Preliminary results," 2014. doi: 10.1177/1932296814525696.

[12] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," BMC Med Res Methodol, vol. 18, no. 1, 2018, doi: 10.1186/s12874-018-0611-x.

[13] S. Kim and D. Kim, "Analyzing mobile application logs using process mining techniques: An application to online bookstores," ICIC Express Letters, Part B: Applications, vol. 9, no. 6, pp. 607–614, 2018, doi: 10.24507/icicelb.09.06.607.

[14] S. Miller et al., "A framework for analyzing and measuring usage and engagement data (AMUSED) in digital interventions: Viewpoint," J Med Internet Res, 2019, doi: 10.2196/10966.

[15] S. Raj, K. Toporski, A. Garrity, J. M. Lee, and M. W. Newman, "'My blood sugar is higher on the weekends': Finding a role for context and context-awareness in the design of health self-management technology," 2019. doi: 10.1145/3290605.3300349.

[16] R. Pawson and N. Tilley, Realistic Evaluation. London: SAGE Publications Ltd, 1997.

[17] M. a F. Kazi and I. Rostila, "The Practice of realist evaluation in two countries," Proceedings of the European Evaluation Society Conference, 2002.

[18] M. J. Grant and A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," Health Info Libr J, 2009, doi: 10.1111/j.1471-1842.2009.00848.x.

[19] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," International Journal of Social Research Methodology: Theory and Practice, 2005, doi: 10.1080/1364557032000119616.

[20] A. C. Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," Ann Intern Med, vol. 169, no. 7, 2018, doi: 10.7326/M18-0850.

[21] A. W. Harzing, "Publish or Perish." 2007. [Online]. Available: https://harzing.com/resources/publish-or-perish

[22] Dina, "How is Scopus Different from PubMed?," MSK Library Blog, 2015. https://library.mskcc.org/blog/2015/12/how-is-scopus-different-than-pubmed/ (accessed Aug. 10, 2021).

[23] Atlas.ti, "Atlast.ti Qualitative Data Analysis." https://atlasti.com/ (accessed Oct. 20, 2020).

[24] C. Zhang and M. Chu, "A research of adoption process of mobile app based on massive log data," 2018. doi: 10.1145/3271972.3271975.

[25] Q. Pham et al., "An analytics platform to evaluate effective engagement with pediatric mobile health apps: Design, development, and formative evaluation," JMIR Mhealth Uhealth, 2018, doi: 10.2196/11447.

[26] O. Andrei, M. Calder, M. Chalmers, A. Morrison, and M. Rost, "Probabilistic formal analysis of app usage to inform redesign," 2016. doi: 10.1007/978-3-319-33693-0_8.

[27] A. Kim, J. Lee, and M. Kim, "Context-Aware Recommendation Model based on Mobile Application Analysis Platform," Multimed Tools Appl, vol. 75, no. 22, 2016, doi: 10.1007/s11042-015-3146-8.

[28] M. Kim, J. Kim, J. Choi, and M. Trivedi, "Mobile Shopping Through Applications: Understanding Application Possession and Mobile Purchase," Journal of Interactive Marketing, vol. 39, 2017, doi: 10.1016/j.intmar.2017.02.001.

[29] J. Y. Soh et al., "Variability in doctors' usage paths of mobile electronic health records across specialties: Comprehensive analysis of log data," JMIR Mhealth Uhealth, 2019, doi: 10.2196/12041.

[30] M. Cheung, J. She, and R. Lam, "Analyzing the user inactiveness in a mobile social game," 2014. doi: 10.1109/iThings.2014.75.

[31] C. Kao, "How product analytics helped us reduce mobile app churn: a case study," Mixpanel, 2020. https://mixpanel.com/blog/how-user-analytics-helped-us-reduce-mobile-app-churn-a-case-study/ (accessed Apr. 20, 2021).

[32] X. Ferre, E. Villalba, H. Julio, and H. Zhu, "Extending mobile app analytics for usability test logging," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, vol. 10515 LNCS. doi: 10.1007/978-3-319-67687-6_9.

[33] Lancaster University, "5 Evaluation perspectives," Evaluation Capacity Building in Widening Participation Practice, 2008. https://www.lancaster.ac.uk/fass/events/capacitybuilding/toolkit/perspectives.htm (accessed May 17, 2021).

[34] A. van Schalkwyk, S. Grobbelaar, and M. Herselman, "A scoping review of the use of data analytics for the evaluation of mhealth applications," 2020.

[35] N. B. Baltierra, K. E. Muessig, E. C. Pike, S. LeGrand, S. S. Bull, and L. B. Hightow-Weidman, "More than just tracking time: Complex measures of user engagement with an internet-based health promotion intervention," J Biomed Inform, vol. 59, 2016, doi: 10.1016/j.jbi.2015.12.015.

[36] Y. Lee, Y. R. Park, J. Kim, J. H. Kim, W. S. Kim, and J. H. Lee, "Usage pattern differences and similarities of mobile electronic medical records among health care providers," JMIR Mhealth Uhealth, vol. 5, no. 12, 2017, doi: 10.2196/mhealth.8855.

[37] S. J. Kim, R. J. H. Wang, and E. C. Malthouse, "The Effects of Adopting and Using a Brand's Mobile Application on Customers' Subsequent Purchase Behavior," Journal of Interactive Marketing, vol. 31, 2015, doi: 10.1016/j.intmar.2015.05.004.

[38] J. E. Owen, B. K. Jaworski, E. Kuhn, K. N. Makin-Byrd, K. M. Ramsey, and J. E. Hoffman, "mHealth in the Wild: Using Novel Data to Examine the Reach, Use, and Impact of PTSD Coach," JMIR Ment Health, 2015, doi: 10.2196/mental.3935.

[39] S. M. Jansen-Kosterink, J. Bergsma, A. Francissen, and A. Naafs, "The first evaluation of a Mobile application to encourage social participation for community-dwelling older adults," Health Technol (Berl), vol. 10, no. 5, 2020, doi: 10.1007/s12553-020-00430-9.

[40] J. E. W. C. van Gemert-Pijnen, S. M. Kelders, and E. T. Bohlmeijer, "Understanding the usage of content in a mental health intervention for depression: An analysis of log data," J Med Internet Res, 2014, doi: 10.2196/jmir.2991.

[41] N. de Cock et al., "Feasibility and impact study of a reward-based mobile application to improve adolescents' snacking habits," Int J Middle East Stud, vol. 21, no. 12, 2018, doi: 10.1017/S1368980018000678.

[42] A. P. Rodriguez Müller, A. Lerusse, T. Steen, and S. van de Walle, "Understanding channel choice in users' reporting behavior: Evidence from a smart mobility case," Gov Inf Q, vol. 38, no. 1, 2021, doi: 10.1016/j.giq.2020.101540.

[43] A. Torbjørnsen, M. C. Småstuen, A. K. Jenum, E. Årsand, and L. Ribu, "Acceptability of an mhealth app intervention for persons with type 2 diabetes and its associations with initial self-management: Randomized controlled trial," JMIR Mhealth Uhealth, 2018, doi: 10.2196/mhealth.8824.

[44] M. Billmann, M. Böhm, and H. Krcmar, "Use of workplace health promotion apps: Analysis of employee log data," Health Policy Technol, vol. 9, no. 3, 2020, doi: 10.1016/j.hlpt.2020.06.003.

[45] Q. A. Rahman et al., "Patterns of User Engagement With the Mobile App, Manage My Pain: Results of a Data Mining Investigation," JMIR Mhealth Uhealth, 2017, doi: 10.2196/mhealth.7871.

[46] S. Kim, K.-H. Lee, H. Hwang, and S. Yoo, "Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital," BMC Med Inform Decis Mak, vol. 16, no. 1, 2015, doi: 10.1186/s12911-016-0249-8.

[47] C. H. Ganoe, H. R. Robinson, M. A. Horning, X. Xie, and J. M. Carroll, "Mobile awareness and participation in community-oriented activities," 2010. doi: 10.1145/1823854.1823868.

[48] C. Chen et al., "Making sense of mobile health data: An open architecture to improve individual- and population-level health," J Med Internet Res, vol. 14, no. 4, 2012, doi: 10.2196/jmir.2152.

[49] J. Y. Park et al., "Lessons learned from the development of health applications in a tertiary hospital," Telemedicine and e-Health, vol. 20, no. 3, 2014, doi: 10.1089/tmj.2013.0192.

[50] P. Saikia, M. Cheung, J. She, and S. Park, "Effectiveness of mobile notification delivery," 2017. doi: 10.1109/MDM.2017.14.

[51] A. T. Chen, S. Wu, K. N. Tomasino, E. G. Lattie, and D. C. Mohr, "A multi-faceted approach to characterizing user behavior and experience in a digital mental health intervention," J Biomed Inform, 2019, doi: 10.1016/j.jbi.2019.103187.

[52] J. Wang et al., "Supporting Caregivers of Children With Acute Lymphoblastic Leukemia via a Smartphone App," CIN: Computers, Informatics, Nursing, vol. 34, no. 11, 2016, doi: 10.1097/cin.0000000000000265.

[53] L. Goutas, B. Hess, and J. Sutanto, "If erring is human, is system use divine? Omission errors during post-adoptive system use," Decis Support Syst, vol. 130, 2020, doi: 10.1016/j.dss.2019.113225.

[54] Y. R. Park et al., "Managing patient-generated health data through mobile personal health records: Analysis of usage data," JMIR Mhealth Uhealth, 2018, doi: 10.2196/mhealth.9620.

[55] K. Kaneko and J. Wang, "Behavior Analysis of Learning Methods to Overcome Stress or Pressure on Learning Situation," 2018. doi: 10.1109/IIAI-AAI.2018.00056.

[56] L. Kooij, P. J. E. Vos, A. Dijkstra, and W. H. van Harten, "Effectiveness of a mobile health and self-management app for high-risk patients with chronic obstructive pulmonary disease in daily clinical practice: Mixed methods evaluation study," JMIR Mhealth Uhealth, vol. 9, no. 2, 2021, doi: 10.2196/21977.

[57] N. Bricon-Souf and C. R. Newman, "Context awareness in health care: A review," Int J Med Inform, vol. 76, no. 1, 2007, doi: 10.1016/j.ijmedinf.2006.01.003.

[58] L. G. Morrison et al., "Comparing usage of a web and app stress management intervention: An observational study," Internet Interv, vol. 12, 2018, doi: 10.1016/j.invent.2018.03.006.

[59] G. Jarvis, "Case study: How Premier Inn uses mobile analytics," Computer Weekly, 2013. https://www.computerweekly.com/news/2240203420/Case-study-How-Premier-Inn-uses-mobile-analytics (accessed Apr. 20, 2021).

[60] G. Jimenez, E. Lum, and J. Car, "Examining diabetes management apps recommended from a Google search: Content analysis," JMIR Mhealth Uhealth, vol. 7, no. 1, 2019, doi: 10.2196/11848.

[61] A. van Schalkwyk, S. Grobbelaar, M. Herselman, and E. Vermeulen, "A conceptual framework for generating insights from log analysis in support of mHealth App evaluations," Master's Thesis, Stellenbosch University, 2022.

Ané van Schalkwyk holds a B.Eng (Industrial) and M. Eng (Industrial) attained both from the University of Stellenbosch, South Africa. She is a Junior Business IT Analyst at Imperial Logistics.

Sara Grobbelaar holds an MPhil in Technology Policy from the University of Cambridge. Sara furthermore holds a B.Eng (electronic), M.Eng (computer) and PhD in (Engineering) form the University of Pretoria, South Africa. She also holds a Post Graduate Diploma in M&E methods from Stellenbosch University, South Africa. She is Professor in the Department of Industrial Engineering at Stellenbosch University. She is also a research associate at the Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University.

Euodia Vermeulen holds a BPhysT, Physiotherapy, B.Honors (Industrial), M. Eng (Industrial) all attained from the University of Pretoria, South Africa and holds a PhD (Industrial Engineering) from Stellenbosch University, South Africa. She is a is a post-doctoral fellow at the Departmental Industrial Engineering at Stellenbosch University.

Marlien Herselman holds a PhD from the University of Pretoria, South Africa and is a rated SA researcher (NRF rating at level C). She is a Chief Researcher at the CSIR in Pretoria, South Africa. In this capacity she is working in the Technology Implementation Monitoring and Evaluation (TIME) research group. She was also Chairperson of the board of Living Labs in Southern Africa network (LLiSA) since its initiation over the 2009-2014 period. She an adjunct professor at the University of Pretoria (Faculty of Education), University of Fort Hare (Information Systems Department), Rhodes University (Faculty of IT) and UNISA (School of Computing). In 2009 she won the IEEE, Women in Engineering: Female scientist of the year award (South African section) and in 2016 the CSIR excellence award for Human Capacity Development.