Text-mining analysis of mHealth research

Bunyamin Ozaydin[1], **Ferhat Zengul**[1], **Nurettin Oner**[1], **Dursun Delen**[2]

[1]Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, USA; [2]Department of Management Science and Information Systems, Center for Health Systems Innovation, Oklahoma State University, Stillwater, OK, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: B Ozaydin, F Zengul, N Oner; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Bunyamin Ozaydin. Department of Health Services Administration, University of Alabama at Birmingham, SHPB 590H, 1720 2nd Ave S, Birmingham, AL 35294-1212, USA. Email: bozaydin@uab.edu.

Abstract: In recent years, because of the advancements in communication and networking technologies, mobile technologies have been developing at an unprecedented rate. mHealth, the use of mobile technologies in medicine, and the related research has also surged parallel to these technological advancements. Although there have been several attempts to review mHealth research through manual processes such as systematic reviews, the sheer magnitude of the number of studies published in recent years makes this task very challenging. The most recent developments in machine learning and text mining offer some potential solutions to address this challenge by allowing analyses of large volumes of texts through semi-automated processes. The objective of this study is to analyze the evolution of mHealth research by utilizing text-mining and natural language processing (NLP) analyses. The study sample included abstracts of 5,644 mHealth research articles, which were gathered from five academic search engines by using search terms such as mobile health, and mHealth. The analysis used the Text Explorer module of JMP Pro 13 and an iterative semi-automated process involving tokenizing, phrasing, and terming. After developing the document term matrix (DTM) analyses such as single value decomposition (SVD), topic, and hierarchical document clustering were performed, along with the topic-informed document clustering approach. The results were presented in the form of word-clouds and trend analyses. There were several major findings regarding research clusters and trends. First, our results confirmed time-dependent nature of terminology use in mHealth research. For example, in earlier versus recent years the use of terminology changed from “mobile phone” to “smartphone” and from “applications” to “apps”. Second, ten clusters for mHealth research were identified including (I) Clinical Research on Lifestyle Management, (II) Community Health, (III) Literature Review, (IV) Medical Interventions, (V) Research Design, (VI) Infrastructure, (VII) Applications, (VIII) Research and Innovation in Health Technologies, (IX) Sensor-based Devices and Measurement Algorithms, (X) Survey-based Research. Third, the trend analyses indicated the infrastructure cluster as the highest percentage researched area until 2014. The Research and Innovation in Health Technologies cluster experienced the largest increase in numbers of publications in recent years, especially after 2014. This study is unique because it is the only known study utilizing text-mining analyses to reveal the streams and trends for mHealth research. The fast growth in mobile technologies is expected to lead to higher numbers of studies focusing on mHealth and its implications for various healthcare outcomes. Findings of this study can be utilized by researchers in identifying areas for future studies.

Keywords: Text mining; literature review; mHealth

Received: 23 November 2017; Accepted: 23 November 2017; Published: 27 December 2017.
doi: 10.21037/mhealth.2017.12.02
View this article at: http://dx.doi.org/10.21037/mhealth.2017.12.02

Introduction

Decades of technological developments, especially within the last ten years, brought the term mobile health or mHealth into the healthcare domain. mHealth refers to the use of mobile technologies such as smart phones and tablets to provide medical or health-related services. These technologies work through mobile networks and interfaces such as apps that not only enable communication between clinicians and patients but also provide a platform to the patients for self-care (1,2). Identification of mHealth as the long awaited technological panacea for challenges of U.S. healthcare by the former U.S. Secretary of Health and Human Services Kathleen Sibelius (3) may be understated when one considers the unprecedented growth rate of smart phones and internet usage in the emerging and developing countries. According to the PEW Research Centers' report, compared to 87% [2015] in the 11 advanced economies, the use of the internet at least occasionally or owning a smart phone in the emerging and developing countries jumped from 45% in 2013 to 54% in 2015 (4). Therefore, evidenced by this growth rate, mHealth may have the potential to be a technological panacea for many challenges of healthcare accessibility around the world.

Along with the mobile technologies, the research on mHealth is also growing and evolving rapidly (5). As evidenced by multiple systematic reviews, mHealth has been explored as a remedy to address various infrastructure, geographic, or disease-specific challenges such as public health surveillance in Sub-Saharan Africa (6), healthy aging in developed countries (7), health promotion and primary prevention among older adults (8), maternal health in low income countries (9), management of tuberculosis, HIV/AIDS, and chronic diseases (10,11), interventions for heart-failure, cardiovascular health care, oncology, and medication adherence (12-15).

Besides the aforementioned systematic reviews that focus on implications of mHealth on a particular disease, patient population, or geographical location, there are also systematic reviews that attempt to summarize the overall mHealth research (5) or develop an ontology for mHealth (16). For example, Ali *et al.* (2016) started their systematic review with 3277 articles indexed in PubMed and subsequently reviewed 515 articles that met their inclusion criteria. Their review highlighted the evolution of mHealth research resulting from changes in the deployment of mobile technologies from PDAs (before 2007) to mobile phones [2007 to 2012] and more recently [2013 and 2014] to smartphones and tablets. In another study, Cameron *et al.* [2017] developed an ontological framework for mHealth and applied the framework into 364 articles from 2014 by coding them into the five ontological dimensions including structure, function, semiotics, stakeholders, and outcomes.

Although existing review studies provide some information about the evolution of mHealth research, they either are too specific to a patient population, disease, and geographical location or rely upon manual categorization of large numbers of studies into subject areas, themes, and ontologies. Moreover, categorizing large numbers of disparate studies from a rapidly growing subject area such as mHealth into systematic reviews and interpreting the results in a logical way can be very challenging. It is also possible that manual categorization of studies into major subject areas or ontologies could be prone to various biases. However, recent developments in data mining technologies, particularly text mining techniques, can address these inherent problems in big and unstructured documents by providing means to reveal underlying patterns and trends through automated or semi-automated categorizations (17). Therefore, the overall aim of this study is to reveal the major subject areas of mHealth research and summarize the evolution and trends within the last ten years.

In this study, building on Delen and Crossland's [2008] work, we utilized semi-automated text categorizations to classify research abstracts into major subject categories, and clusters. We also took publication years into consideration to reveal the research trends over time for each major subject category.

Methods

This study loosely follows the text mining process that is a derivative of Cross-Industry Standard Process for Data Mining (CRISP-DM) as described by Miner *et al.* (18). CRISP-DM is an iterative process, consisting of the following steps: (I) determining the purpose of the study, (II) exploring the availability and the nature of the data, (III) preparing the data, (IV) deploying and assessing the model, (V) evaluating the findings, and (VI) deploying the results. Depending on the project, steps (III) and (IV) of CRISP-DM would have significant differences, hence the rest of the methods section will focus there.

Phase 1: document search and establishment of corpus

In text mining, a corpus refers to a collection of documents

Table 1 Database search options

Database	Search in	Publication type	Language
SCOPUS	Title, abstract, keyword	Article, Conference Paper, Review, Book Chapter, Article in Press, Conference Review	English
PubMed	Title/abstract	Journal article, congresses	English
IEEE Xplore	Metadata only	Conference Publications, Journals & Magazines, Early Access Articles, Books & eBooks	English
ABI/Inform	Anywhere except full text	Article, Conference, Conference Paper, Conference Proceeding, Feature, General information, Literature Review, Report, Review (in Conference Papers & Proceedings, Scholarly Journals, Books, Working Papers)	English
ACM-DL	Title, abstract, author keyword	Not applicable	Not applicable

to be analyzed (18). To create a corpus for this work, we identified five databases to search for mHealth research articles: SCOPUS, PubMed, IEEE Xplore, ABI/Inform, and ACM-DL. SCOPUS returned the largest number of publications, followed by PubMed; however, we included the other databases to be able to capture mHealth research produced by non-healthcare disciplines as well. Databases have different search engine options; *Table 1* summarizes advanced search options used for each.

In PubMed, searching MeSH terms was considered but declined because MeSH terms 'mHealth' and 'telemedicine' belong to the same concept, and abstract search already included author keywords (19,20). Inclusion of other publication types in PubMed did not make a significant difference. In IEEE Xplore, searching with the "metadata only" option includes searches in abstract, title text, and indexing terms, and metadata is a combined field that allows searching the Author Keywords, IEEE Terms, INSPEC Terms, and Mesh Terms (21). Therefore, in IEEE Xplore, we sufficed searching only in metadata.

We initially performed a pilot search in SCOPUS, which returned 7,396 records for the original query. We realized that "Mobile Health Unit" was frequently used as a keyword in earlier years. Out of 2,304 records that had this keyword, only 556 were published after 2008. In contrast, when "mHealth" was searched alone, out of 2,622 records, only 13 were published before 2008. Manual examination of about 50 of the 2,304 records also indicated that they were not related to mHealth. Therefore, we decided to change the query to be: ("mHealth" OR "m-Health" OR "mobile health" OR "mobile-health") AND NOT "Mobile Health Units". This modified search criterion was used for all databases resulting in a total of 10,079 records. Distribution of these records per database is shown in *Figure 1*. The results of each database search were exported in RIS format (including abstracts) to be imported into EndNote reference management tool. ACM-DL had only a plain text file export option for EndNote that could not export the abstracts.

Duplicate document identification and exclusions. Once all records were imported and merged in EndNote, we considered using the duplicate identification function of EndNote to exclude duplicate records. However, we concluded that manual review and removal was not feasible due to the large number of records, and EndNote's automatic duplicate removal is not flexible enough. We wanted to control which data elements were considered for duplicate identification and to remove the duplicate records from the database with the smaller number of records, so that we could keep the records as consistent as possible. Therefore, we exported merged records from EndNote, imported them into Excel, and assigned each record a unique ID number. We created a new EndNote output style based on EndNote's tab-delimited style and modified data to reformat author delimiters, so that data could be appropriately exported into Excel (22). Once the "Reference Type", "Type of Work", "Author", "Year", "Title", "Keywords", "Abstract", "Source Database", and "Secondary Title" (storing journal name) data elements were imported into Excel, we initially considered identifying duplicates based on both title and year; however, using only title was found to be more reliable, as different databases may report the same record in different years. A total of 3,703 duplicate records were removed by using the following steps: (I) 36 records that had the same title 4 to 8 times were manually examined and removed; (II) 596 duplicate records from ACM-DL were removed since ACM-DL could not export abstracts; (III) 88 duplicate records from ABI/Inform were removed; (IV) 6 duplicate records manually identified as lacking an abstract were removed from SCOPUS,

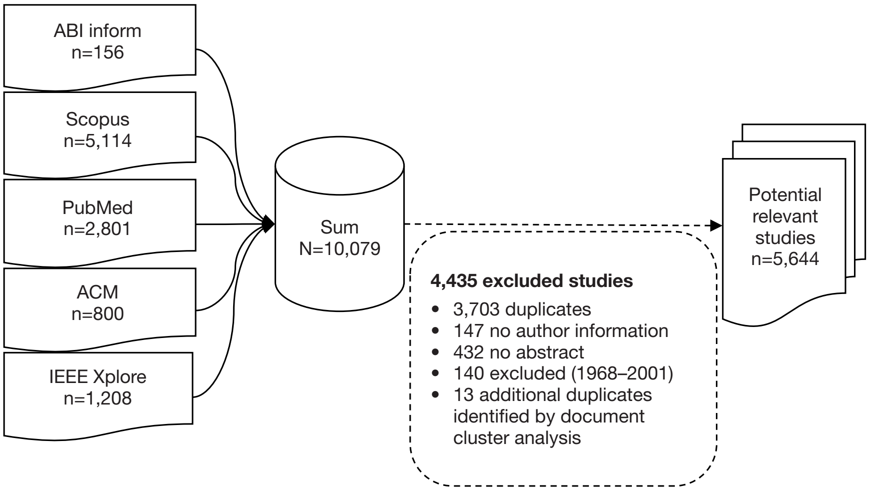
ABI inform
n=156
Scopus
n=5,114
PubMed
n=2,801
ACM
n=800
IEEE Xplore
n=1,208
Sum
N=10,079
Potential relevant studies
n=5,644
4,435 excluded studies
- 3,703 duplicates
- 147 no author information
- 432 no abstract
- 140 excluded (1968–2001)
- 13 additional duplicates identified by document cluster analysis

Figure 1 Flow diagram of included studies in text mining.

because their duplicates from another source included an abstract; (V) 919 and 1,867 duplicate records were removed from IEEE Xplore and PubMed, respectively; (VI) The remaining 376 duplicate/triplicate records were manually reviewed and 191 were removed, preferring journal publications over conference publications, and records with fewer missing data elements or more accurate information.

Furthermore, records with missing author information were examined and 147 were removed for one of three reasons: (I) 83 had "Type of Work" as "Conference Review"; (II) 34 had "Reference Type" as "Conference Proceedings" or "Book/book section" and were referred as title, table of contents, index, appendix, or introduction pages; and (III) 30 were journal articles with missing author information.

Title, abstract, and keywords data elements were considered for inclusion in the text mining analysis. However, only abstract was included, because the abstract usually includes all terms in the title and keywords. Therefore, from the remaining 6,229 records, 432 (191 ACM-DL, 142 SCOPUS, 93 PubMed, 6 ABI/Inform) records that were missing abstract information were removed. Based on reading the abstracts, we recognized that the "mobile health" concept in older articles was used in the context of delivering healthcare in a mobile setting, such as a specially-fitted vehicle, and not in the current context of mobile technologies. Considering the use of mHealth concept today and manually examining a sample of older articles, 140 records published from 1968 to 2001 were also excluded, bringing the data set to 5,657 records (4,668 SCOPUS, 818 PubMed, 254 IEEE Xplore, 57 ABI/Inform) to be included in the analysis. The process of elimination of duplicate documents and other exclusions is summarized in *Figure 1*.

Phase 2: generating and curating the terms list

Text mining was performed using the Text Explorer module of JMP Pro (23). Throughout this paper, whenever a particular function of JMP Text Explorer is used, we denoted it by capitalizing the first letter of each word.

The Excel file obtained at the end of document exclusion process was imported into JMP and an iterative process of curation and analysis (each informing the other) of a terms list was performed as described in JMP documentation (24). The curation process of terms list itself is iterative and summarized in *Figure 2*.

The terminology used to describe these text mining processes is provided in *Table 2*, which along with description of a particular text mining process, illustrates its application in this study by showing the effects of the process on an example phrase or by describing study-specific use. During tokenizing, the entire corpus is converted into lower-case, tokenizing rules are applied to break text into tokens and determine which tokens are included for each record, and re-coding rules are applied to the tokens list to group indicated tokens together. During phrasing, the entire corpus is searched for number of tokens that are used as a group. During terming, JMP Text Explorer creates an initial terms list from tokens along with their frequencies by excluding tokens based on minimum

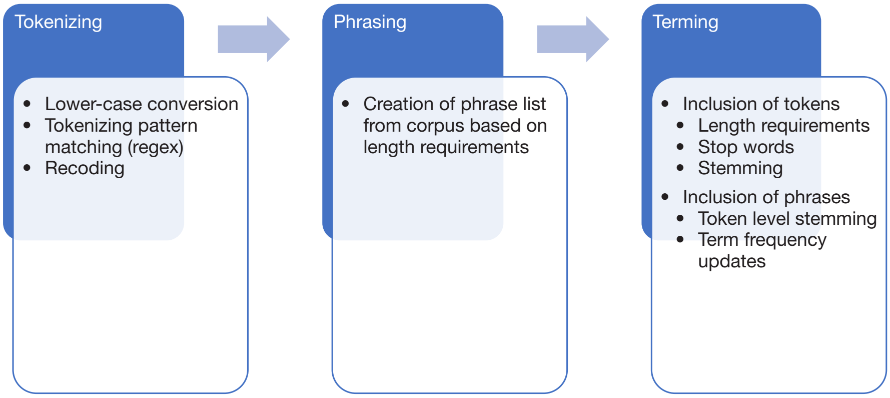

Figure 2 The iterative process of curating the terms list.

Table 2 Terminology used for describing the term generation and curation process

Concept	Description	Application/example text*
Token	The smallest unit of text (group of characters) corresponding to a concept, like a word in a given text	Short, messaging, service, sms, is, frequently, used, in, mobile, health, applications
Tokenization	Breaking text into tokens	
Regular expression (regex)	Language that describes patterns and rules used to describe matching text during tokenization	Adding 'dash' to the list of characters to match words with those embedded characters allows mobile-health to become a single token: [&'-]
Stop word	Tokens excluded from analysis	is, in
Recoding	Renaming tokens in order to group or ungroup them. Frequently used to indicate synonyms	mobile-health → mHealth, m-health → mHealth
Stemming/ lemmatization	Reduction of tokens into their simplest form	(message, messaged, messages, messaging) → messag·
Phrase	Combination of a small number of tokens	short· messag· servic·
Term	A token or a phrase	short· messag· servic·, sms, frequently, used, mhealth, applications
Document	The unstructured text included in the analysis for a particular record	The abstract of a particular publication
Corpus	The collection of documents included in the analysis	All of the abstracts included in the analysis
Document term matrix (DTM)	The matrix where rows correspond to document, columns correspond to terms and each cell corresponds to values of analysis based on the weighing option	Frequencies or TF/IDF values for each document/term pair

*Short messaging service (SMS) is frequently used in mobile-health applications.

and maximum length limits and stop words and applying stemming rules. Additionally, it displays a phrase list along with their frequencies, allowing users to add phrases from it into the terms list. As a phrase is added to the terms list, stemming rules are applied to the tokens that make up the phrase and the token term frequencies are updated to avoid duplicate counting.

This study included only the abstracts to establish the corpus, considering each publication's abstract as a document. In JMP Text Explorer, the following are identified as optimal by trying various options and reviewing the resulting term and phrase lists: English for language, “Stem for Combining” for stemming, custom regex (Regular Expression) for tokenizing, 3 to 50 for the range of number of characters per token, and the default 4 for the maximum number of words per phrase. We found the default regex options in the JMP Text Explorer to be very effective and used them as a basis: all default regex syntax, except the “Words” pattern, which were kept as is and their results were set to be ignored. The “Words” regex syntax was modified to include hyphen as an allowed embedded character, when used in between two regular words: \[([\□]{1,99}(?:[&'-.]{0,1}[0-9\□]{1,99}){0,99})]\.

The Text Explorer user interface, which displays dynamic lists of terms and identified phrases and allows users to indicate stop words, phrases to be included in the terms list, token level recoding and stemming rules, and exceptions for built-in stems, phrases, and stop words, was used to perform the terms list curation process. The following actions, taken to finalize the curation process, were implemented iteratively as each decision made at a given step affects other parts of both the curation and the analysis: (I) We manually reviewed every phrase with a frequency of at least 10 and included those that made contextual sense for this study. We also identified phrase exceptions, to avoid double counting a phrase and its abbreviation. For example, since ‘electronic health record’ was added to the terms list as a phrase, ‘electronic health record ehr’ was added to the phrase exception list. (II) We manually reviewed every term with a frequency of at least 18, and added tokens into the stop words list not contextually valuable for our study. These stop words included verbs, prepositions, conjunctions, journal abbreviations, etc. We added stem exceptions to differentiate tokens. For example, to ensure ‘aids’ was not stemmed into ‘aid,’ or ‘careful’ was not stemmed into ‘care.’ We also re-coded some tokens to group them together and to be able to indicate synonyms.

Phase 3: analyzing the terms list

In the analyses stage, a bag of words approach based on the term counts was utilized. After achieving the curated term list through tokenizing, phasing, and terming (*Figure 2*), analyses were performed on the Document Term Matrix (DTM) based upon the curated terms list. In the DTM, rows represent the documents, columns represent the terms, and each cell represents a value of analysis based on the weighting option. JMP provides five different weighting options: binary, ternary, frequency, log frequency, and TF/IDF (term frequency into inverse document). We used TF/IDF, the most commonly used weighting option, since it normalizes the raw indices and reduces the potential bias that may arise when only frequencies are used (17,18). Due to DTM’s large and rather sparse (many zeroes) nature, dimension reduction in DTM is a necessary step before attempting to reveal existing patterns in the unstructured data. We utilized Latent Semantic Analysis (LSA), a technique similar to principle component analysis that focuses on dimension reduction through Singular Value Decomposition (SVD). SVD represents the matrix as a series of linear approximations to reveal underlying meanings (17). LSA captures connections among different words with similar meanings or topic areas (24). During this step, we determined the number of singular vectors as 14 and minimum term frequencies as 18 through iterative processes. To achieve our main goal of identifying document clusters indicative of the underlying research themes by using TF/IDF values of the terms they contain, we utilized several analyses.

First, we deployed the straightforward Cluster Documents method, a hierarchical clustering of the documents. However, despite our multiple and iterative efforts, term groups for the generated document clusters were not distinct enough, making it difficult to name the resulting document clusters. There is a feature in other text mining tools, such as SAS Text Miner, that displays term groups for each document cluster automatically. However, we had to do this task manually in JMP Text Explorer by saving TF/IDF values into the data table in DTM format, along with the cluster assignment for each document. We then used this information to generate term groups for each document cluster by selecting 30 terms with highest aggregated TF/IDF values within each cluster.

Although Cluster Documents analysis had limitations, we found it to be useful in identifying outliers by examining the clusters with a proportionally very small number of documents. This led us to identify a few documents that were not related to mHealth and a few duplicates that were missed in earlier stages due to minor spelling variations in their titles. As a result, we excluded 13 additional documents that reduced our final sample to 5,644 documents. Secondly, the limitation in Cluster Documents analysis led us to use Topic Analysis (TA), an option similar to factor analysis that performs orthogonal varimax rotation on the SVD of the DTM (24). The results of TA are displayed on the Topic Words report, which exhibited 15 to 20 terms with

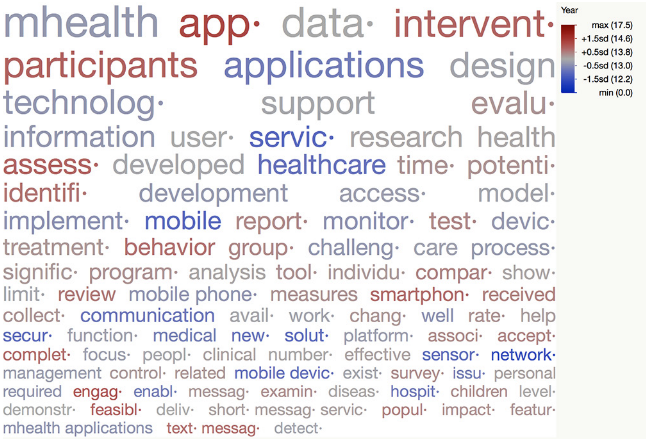

Figure 3 Word cloud of the entire corpus based on frequency emphasizing changes in use of terms over the years.

the highest topic scores in each topic. Again, through iterative review of the Topic Words report, we determined the numbers of topics as ten. Since our focus is on the document clusters, not the topics, we used the results of TA to generate topic-informed document clustering. JMP Text Explorer does not have a feature to save the assigned topic-informed document cluster for each document the way it saves the assigned document cluster. Therefore, we used Save Document Topic Vectors to save the topic scores for each document, based on which we calculated the assigned topic-informed document clusters. Then, we used the same method described in the previous paragraph to generate term groups of each topic-informed document cluster. Finally, we exported thirty terms with the highest TF/IDF scores from each of the ten clusters into R statistical package to create their word clouds.

Results

We used the Display Word Cloud option in JMP Text Explorer to generate *Figure 3*, which exhibits the word cloud of the entire corpus based on the frequency of terms and the changes in the use of terms over years. A term's font size is larger as its frequency is higher. The terms with higher frequencies in the earlier years are shown in blue colors, whereas the ones with higher frequencies in recent years are shown in red colors. For example, the blue color of the term "applications" indicates more frequent usage of the term earlier, whereas the red color of "app." indicates more dominant usage of it in recent years.

Figure 4 exhibits the word clouds for topic-informed document clusters. These clusters were named by five researchers including the four authors and an expert in mHealth through several rounds of an iterative and blinded process. In the first round, each of the five researchers individually named the word clouds in a Microsoft Excel file then shared the file with a third-party individual. The third-party individual combined and scrambled the five different names for each cluster and shared the result with the five researchers without revealing the identity of the persons who suggested the names. During the second round, the five researchers renamed the clusters by either using an existing name or determining a new name. In the third round, researchers discussed each of five name suggestions for each clusters and determined the final names for each of these clusters.

Cluster 1 (C-1): clinical research and lifestyle management

The most frequently appeared terms in C-1 includes: participants, intervention, and physical activity. Clinical research was identified as one of themes for this cluster of studies based on the frequently appearing terms such as participants, intervention, control group, intervention group, and trial. Lifestyle management was identified as another important theme based on the frequently appearing terms such as physical activity, behaviors, exercise, diet, adherence, weight loss, body mass index, and obesity. Overall, the terms in this cluster suggest that there is a stream of mHealth research focusing on clinical areas and investigating implications of mHealth on lifestyle management choices and activities.

Cluster 2 (C-2): community health

The most frequently appeared terms in C-2 includes: children, women, and short-messaging services. Some other frequently appearing words such as community vaccination, intervention, services, pregnancy, health worker, maternity, rural, HIV, children and women services (chws), care, access, and clinic suggest a cluster of research around community and public health. The research in C-2 focuses on investigating the implications of mHealth interventions on various community health related outcomes.

Cluster 3 (C-3): literature review

The most frequently appeared terms in C-3 are search, review, systematic review and intervention. Additionally,

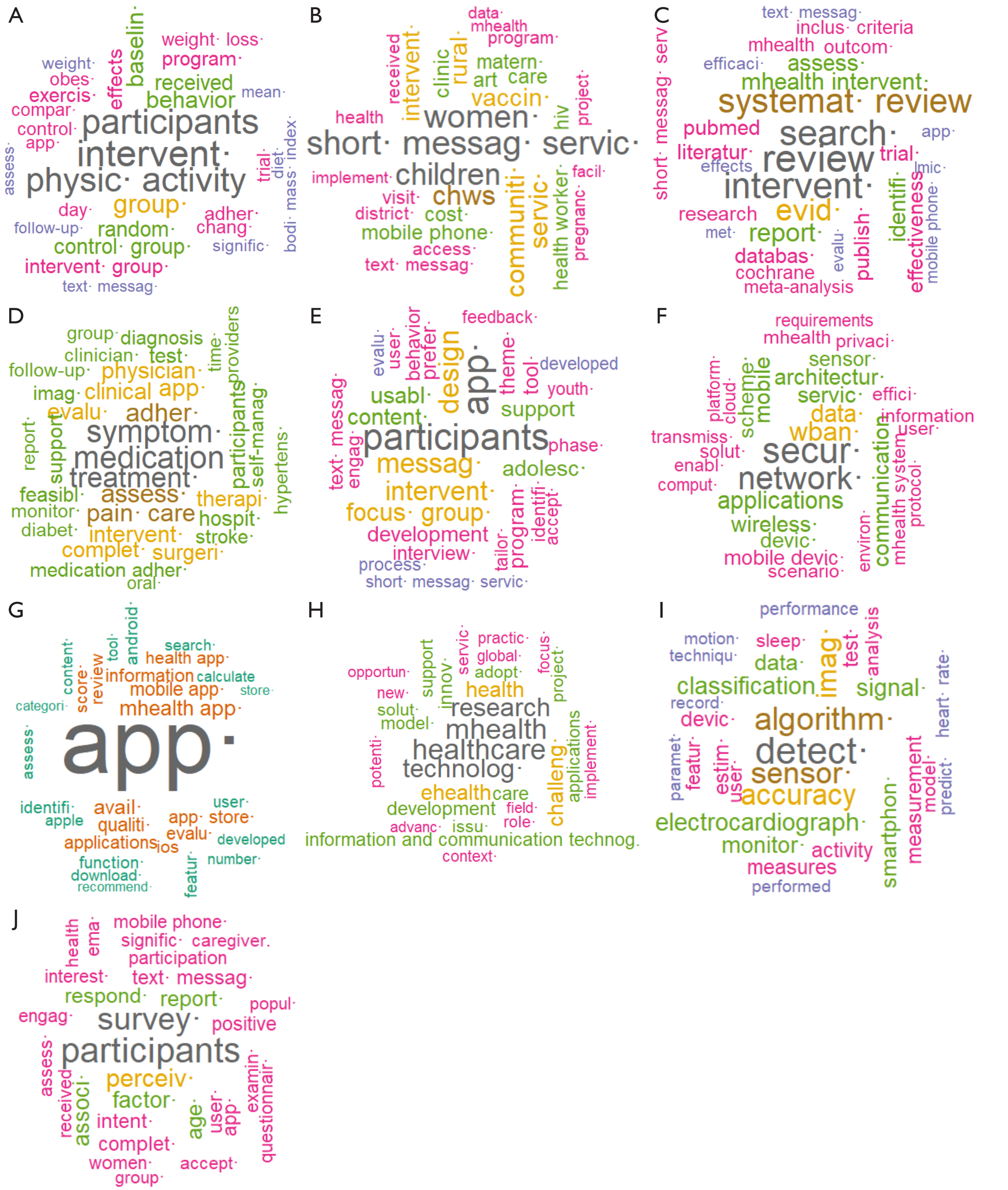

Figure 4 Document clusters based on topic assignments: (A) clinical research on lifestyle management; (B) community health; (C) literature review; (D) medical interventions; (E) research design; (F) infrastructure; (G) applications; (H) research and innovation in health technologies; (I) sensor based devices and measurement algorithms; (J) survey based research.

some other terms such as Cochrane, PubMed, database, meta-analysis and literature suggest that there is a stream of research focusing on reviewing the literature and summarizing implications of mHealth interventions on various healthcare outcomes.

Cluster 4 (C-4): medical interventions

The most frequently surfaced terms in C-4 are symptom, medication, and treatment. These terms combined with terms such as adherence, diagnosis, therapy, follow-up, physician, evaluation, and assess suggest that there is a cluster of research focusing on the use mHealth in various medical interventions and treatments. Some other frequently appeared terms such as hypertension, pain, surgery, and stroke highlight the medical conditions to which mHealth interventions were applied. The terms medication adherence, adherence, and self-management indicate studies examining the effects of mHealth interventions on medication adherence and self-management of medication regimens or on the processes and outcomes in general.

Cluster 5 (C-5): research design

The most frequently surfaced terms in C-5 are participants and apps. These terms taken together with the others such as design, focus group, intervention, content, and messaging suggest that mHealth research in this cluster mainly explores various methods and designs applied in studies. Some other frequent, yet less noticeable, terms such as youth, adolescent, user, program, support, and content specify study participants and potential processes and programs that are used to implement these programs to influence certain behaviors or outcomes.

Cluster 6 (C-6): infrastructure

The most frequently surfaced terms in C-6 are secure and network. Other terms such as applications, wireless, data, communications, wireless body area network (wban), service delivery, cloud, platform, mobile devices, sensor, architecture, and scheme suggest this cluster of mHealth research focusing on infrastructural issues. The frequent use of some other terms such as privacy, enable, solution, scenario, efficient and requirements suggest that the infrastructural issues related to mHealth are generally used within a context of finding solutions, developing scenarios and enabling privacy.

Cluster 7 (C-7): applications

Having the most frequently surfaced term "app", C-7 is the cluster most distinguished from all other clusters. This finding suggests that there is an important cluster of mHealth research exploring various apps and their implications on health. This finding is supported by other less conspicuous terms within C-7 such as mHealth app, health app, download, Android, Apple, app store, features, and functionalities. Some of the action terms such as assess, evaluate, developed, search, score, calculate and categorize indicate that these apps are used as health interventions in various research contexts and their implications are evaluated or assessed.

Cluster 8 (C-8): research and innovation in health technologies

The most frequently surfaced terms in C-8 are mHealth, technology, healthcare, and research. Additionally, terms such as information and communication technologies, applications, and advancement suggest that C-8 focuses on innovative technologies in healthcare research. C-8 is the only cluster that eHealth appeared in, along with mHealth. Other terms such as challenge, development, adoption, implement, focus, context, global, and support indicates the context for the mHealth research for this cluster.

Cluster 9 (C-9): sensor based devices and measurement algorithms

The most frequently appeared terms in C-9 are detect, algorithm, sensor, and accuracy. There are also specific terms indicating device types such as electrocardiograph, monitor, and smartphone. Additional terms such as heart rate, predict, classify, record, estimate, measure, and analyses suggest C-9 focuses on use of sensor based devices, measurement algorithms, and predictive analyses.

Cluster 10 (C-10): survey based research

The most frequently appeared terms in C-10 were survey and participants. Other terms such as questionnaire, respond, participation, complete, age, user, and report suggest that C-10 studies utilize surveys or questionnaires to collect data by targeting various segments of populations

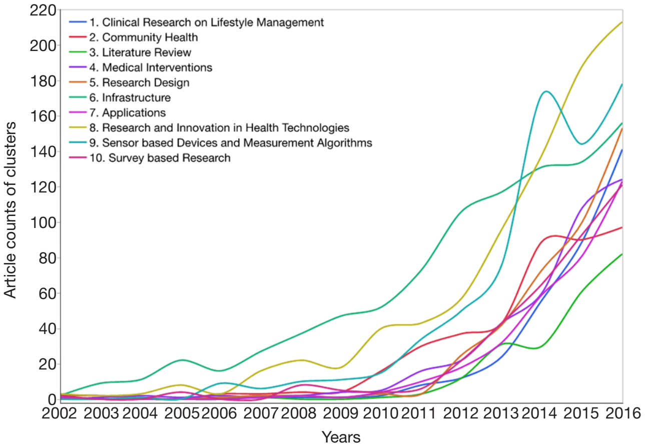

Figure 5 Cluster trends by article count.

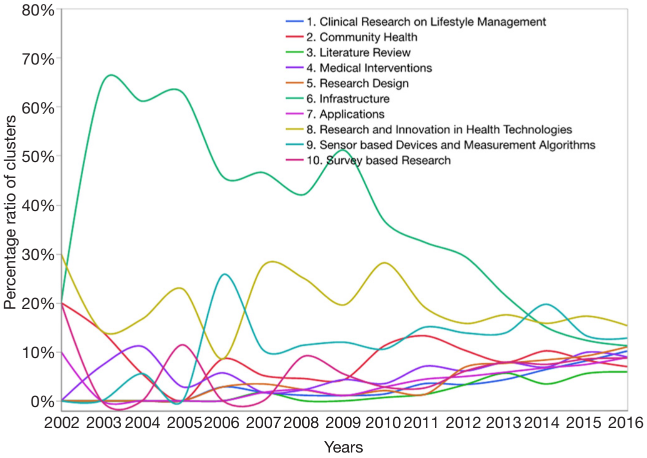

Figure 6 Cluster trends by ratio percentages.

such as caregivers, and women.

Figure 5 exhibits the research trends by number of documents for each of the above-mentioned ten clusters. The increasing research trend for all ten clusters, with the exception of the survey based research cluster (C-10), suggests the momentum that mHealth research has gained in recent years. Among others, the research and innovation in health technologies cluster (C-8) exhibits the most articles with a continuously increasing trend, especially after year 2012. Despite having the highest article counts in 2013 and 2014, the sensor based devices and measurement algorithms cluster(C-9) exhibits a notable dip in 2015.

Figure 6 displays cluster research trends by ratio percentages. The majority of research trends exhibits similar patterns with an exception of the infrastructure cluster (C-6). Even though the infrastructure cluster starts around 20% in 2002 like others, it later exhibits a steep jump to over 65% in 2003, and then displays an intermittent decline.

Discussion

There are several interesting points from our results that emerge when the counts and ratios of each cluster across years in *Figures 5,6* are interpreted along with the findings in the word clouds in *Figure 4*.

First, as observed in *Figures 4-6*, mHealth research focusing on infrastructure (C-6) has been the most investigated cluster starting from 2003 through 2014 in terms of both count and ratio. This was more prevalent during the earlier years; for example, between 2003 and 2006 more than 50% of all mHealth articles were classified under the infrastructure cluster (C-6). Interpretation of the terms in this cluster with its corresponding trends suggests the importance of infrastructure challenges such as availability of WBAN, wireless networks, security and privacy, and cloud platforms for mHealth research. To address infrastructural issues, studies explored various solutions, including extension of the local services offered by a Body Area Network (BAN) (25); providing secure access to electronic health records (EHR) via Regional Health Information Networks (RHINs) (26); and analyzing potential of 3G wireless networks in supporting m-health services (27,28). Obviously, since the introduction of first mobile phone in 1973 (5), infrastructure supporting mobile devices has advanced substantially. Given that the infrastructure cluster represents the backbone of mHealth research, we highly recommend sustained attention by researchers on this cluster.

Second, as it is observed in *Figures 5,6*, there has been an accelerating research trend in the health technologies and innovation cluster (C-8), which is expected given the ever-accelerating developments in technologies that support mobile health innovations. The literature highlights the advancements in database computing, internet, bio-sensing, diagnostic, sensor, wireless, communication, and network technologies (29-33) as the enhancers of mHealth innovations and research. However, one should also consider the importance of diffusion of these technologies (34). As higher percentages of the population adopt and use mobile technologies, the mHealth research potential would also grow. Therefore, we believe that while investigating in the technology cluster, researchers should consider both technological advancement and adoption level.

Third, our text-mining analyses confirm the findings of an earlier systematic review (5) in regards to the evolution of mHealth research mirroring the developments in mobile devices and technologies. For example, *Figure 3* shows that during earlier years of mHealth research, researchers used the terms mobile device or mobile phone, versus the term smartphones, which is used more frequently during more recent years.

Fourth, some of the ten clusters identified in our study exhibit resemblance to five ontological dimensions identified in a study reported earlier in 2017 by Cameron, Ramaprasad, and Syn (16) that utilized 364 articles. These five dimensions were Structure, Function, Semiotics, Stakeholders, and Outcome (16). From these ontological dimensions, the structure—defined as the structural elements of a mHealth system—exhibits a close similarity to our infrastructure cluster (C-6), where the same terms such as network, platform, applications, wireless, and devices were identified in structure/infrastructure groupings of both studies. Cameron *et al.* argued that the structure dimension is biased towards technologies such as applications and fails to address infrastructural issues such as network processes and policies (16). However, we did not observe this bias. On the contrary, in our infrastructure cluster the terms such as network and security were the most frequent. One potential reason for this difference is having separate clusters of application (C-7) and technology (C-8) in our study.

For the other ontological dimensions, it was harder to pinpoint similarities due to the inherent difference between their study and ours. They conceptually developed ontological dimensions and manually coded each study into these dimensions (16). Therefore, the original terms might have been masked during this manual process and coded into another term by the researchers. However, we utilized topic-informed document cluster analysis to reveal the existing patterns without making any terms. We recommend that a future study explore the similarities and differences between these two approaches by utilizing the manual coding sheets from the Cameron *et al.*'s 2014 study and text mining results from our study. A more comprehensive road map for mHealth research can be generated by combining both approaches.

Conclusions

In this study, we explored the evolution of mHealth research by utilizing text-mining analyses of 5,644 manuscript abstracts. The main objective of the study was to understand the mHealth research trends and define a global perspective on the past, present, and future of mHealth research as an academic field. Our findings revealed ten clusters for mHealth research. The evolution of these clusters over the last 12 years suggests that mHealth research is expanding as the various underlying wireless, sensor, network, communication, and internet technologies advance. However, mHealth would be still considered at its infancy as an academic research field given the numbers of publications. More studies and use of combined methods are needed to develop a comprehensive ontological roadmap for mHealth research.

Acknowledgements

We are grateful to Dr. Donna Slovensky for encouraging us to write this article, discussions she engaged with us throughout the preparation, and feedback she provided to the final version of this work.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Malvey D, Slovensky DJ. mHealth. New York: Springer US; 2014.
2. AHIMA. Best Practices for Mobile Health? There's an APP Guide For That. Available online: https://myphr.com/HealthLiteracy/MX7644_myPHRbrochure.final7-3-13.pdf
3. Levy D. Emergin mHealth: Paths for growth. 2012. Available online: https://www.pwc.com/gx/en/healthcare/mhealth/assets/pwc-emerging-mhealth-full.pdf (Accessed May 22, 2017).
4. Poushter J. Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. 2016. Available online: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/ (Accessed May 23, 2017).
5. Ali EE, Chew L, Yap KY. Evolution and current status of mhealth research: a systematic review. BMJ Innovations 2016;2:33-40.
6. Brinkel J, Krämer A, Krumkamp R, et al. Mobile phone-based mHealth approaches for public health surveillance in sub-Saharan Africa: a systematic review. Int J Environ Res Public Health 2014;11:11559-82.

7. Henriquez-Camacho C, Losa J, Miranda JJ, et al. Addressing healthy aging populations in developing countries: Unlocking the opportunity of eHealth and mHealth. Emerg Themes Epidemiol 2014;11:136.
8. Kampmeijer R, Pavlova M, Tambor M, et al. The use of e-health and m-health tools in health promotion and primary prevention among older adults: A systematic literature review. BMC Health Serv Res 2016;16 Suppl 5:290.
9. Colaci D, Chaudhri S, Vasan A. mHealth Interventions in Low-Income Countries to Address Maternal Health: A Systematic Review. Ann Glob Health 2016;82:922-35.
10. Devi BR, Syed-Abdul S, Kumar A, et al. MHealth: An updated systematic review with a focus on HIV/AIDS and tuberculosis long term management using mobile phones. Comput Methods Programs Biomed 2015;122:257-65.
11. Hamine S, Gerth-Guyette E, Faulx D, et al. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: A systematic review. J Med Internet Res 2015;17:e52.
12. Cajita MI, Gleason KT, Han HR. A systematic review of mhealth-based heart failure interventions. J Cardiovasc Nurs 2016;31:E10-22.
13. Chow CK, Ariyarathna N, Islam SM, et al. mHealth in Cardiovascular Health Care. Heart Lung Circ 2016;25:802-7.
14. Brouard B, Bardo P, Vignot M, et al. eHealth and mHealth: current developments in 2014 and perspectives in oncology. Bull Cancer 2014;101:940-50.
15. Dekoekkoek T, Given B, Given CW, et al. mHealth SMS text messaging interventions and to promote medication adherence: An integrative review. J Clin Nurs 2015;24:2722-35.
16. Cameron JD, Ramaprasad A, Syn T. An ontology of and roadmap for mHealth research. Int J Med Inform 2017;100:16-25.
17. Kim YM, Delen D. Medical informatics research trend analysis: A text mining approach. Health Informatics Journal 2016. Available online: https://doi.org/10.1177/1460458216678443
18. Miner G, Elder J, Fast A, et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Cambridge, MA: Academic Press; 2012.
19. Neveol A, Dogan RI, Lu Z. Author keywords in biomedical journal articles. AMIA Annu Symp Proc 2010;2010:537-41.
20. Using Command Search. Available online: http://ieeexplore.ieee.org/Xplorehelp/#/searching-ieee-xplore/command-search#summary-of-data-fields (Accessed Oct 13, 2017).
21. Torre S. Author Keywords in PubMed. NLM Tech Bull 2013;(390):e2.
22. Endnote X7 for Windows: modifying reference types and output styles for Windows. Available online: http://endnote.com/sites/en/files/m/pdf/en-x7-win-editing-reference-types-styles.pdf (Accessed Oct 13, 2017).
23. JMP Pro Text Explorer. Available online: http://www.jmp.com/support/help/13-2/Text_Explorer_Overview.shtml# (Accessed Oct 13, 2017).
24. JMP® 13 Basic Analysis. Cary, NC: SAS Institute; 2016.
25. Dokovsky N, Van Halteren A, Widya I. BANip: Enabling remote healthcare monitoring with body area networks. IEEE Communications Magazine 2006;44:91-6.
26. Orphanoudakis S. HYGEIAnet: the integrated regional health information network of Crete. Stud Health Technol Inform 2004;100:66-78.
27. Bults R, Wac K, Van Halteren A, et al. editors. Goodput analysis of 3G wireless networks supporting m-health services. Zagreb: 8th International Conference on Telecommunications (ConTEL), 2005.
28. Garawi S, Istepanian RS, Abu-Rgheff MA. 3G wireless communications for mobile robotic tele-ultrasonography systems. IEEE Communications Magazine 2006;44:91-6.
29. Yoo S, Kim B, Park H, et al. Realization of real-time clinical data integration using advanced database technology. AMIA Annu Symp Proc 2003:738-42.
30. Roehrs A, Da Costa CA, Da Rosa Righi R, et al. Personal health records: A systematic literature review. J Med Internet Res 2017;19:e13.
31. Quesada-González D, Merkoçi A. Mobile phone-based biosensing: An emerging "diagnostic and communication" technology. Biosens Bioelectron 2017;92:549-62.
32. Istepanian RS. m-Health Computing: m-Health 2.0, Social Networks, Health Apps, Cloud, and Big Health Data. In: m-Health:Fundamentals and Applications. Hoboken, NJ: John Wiley & Sons, Inc., 2016;424.
33. De Rosis S, Vainieri M. Incentivizing ICT in healthcare: A comparative analysis of incentive schemes in Italian Regions. International Journal of Healthcare Management 2017;10:1-12.
34. Rogers EM. Diffusion of innovations. New York: Simon and Schuster; 2010.

doi: 10.21037/mhealth.2017.12.02
Cite this article as: Ozaydin B, Zengul F, Oner N, Delen D. Text-mining analysis of mHealth research. mHealth 2017;3:53.