Smart and Connected Health Projects: Characteristics and Research Challenges

Jiangping Chen[1][0000-0002-7016-4583], Minghong Chen[2][0000-0002-6989-2719], Jingye Qu[3][0000-0002-1715-1919],

Haihua Chen[1][0000-0003-2806-3938], Juncheng Ding[4][0000-0002-9529-6056]

[1] Department of Information Science, University of North Texas, Denton TX 76203, USA
[2] School of Information Management, Sun Yat-sen University, Guangzhou 510006, CHINA
[3]School of Information Technology and Media, Beihua University, Jilin, 132013, CHINA
[4] Department of Computer Science, University of North Texas, Denton TX 76203, USA
Jiangping.Chen@unt.edu

Abstract. The Smart and Connected Health (SCH) program at the National Science Foundation (NSF) has been established as a stand-alone solicitation since 2012. This article reviews and analyzes the 100 projects that have been funded since 2012 to understand their characteristics and the research challenges they have addressed in SCH. Descriptive analysis, topic analysis based on Latent Dirichlet Allocation (LDA), and comparative content analysis were performed. Our study indicated that NSF SCH projects, featured with collaborative and multidisciplinary research endeavor, have been exploring more than 36 diseases or health problems, and five major research challenges including electronic health record (HER) data processing, system design or computational model building, personalized or patient-centered medicine, training and education, and privacy preserving. Much more research projects are needed to investigate algorithms, devices, and impacts of smart health on diseases and communities.

Keywords: Smart and Connected Health, NSF Projects, Text Analysis, Data Analysis.

1 Introduction

With the rapid development of the infrastructures and technologies of smart cities that reconstruct the thinking behind existing healthcare systems and telemedicine, a new and ubiquitous concept called smart health, or smart and connected health (SCH) has emerged [1], leading the innovation of health care service mechanism. Although SCH has not been precisely defined, it refers to any digital healthcare solutions or systems that can operate remotely with integration of innovative computational and engineering approaches to support the transformation of health and medicine services [2][3]. According to Clancy [3], SCH defines not only information communication technology development, but also a state-of-thinking, a way of lifestyle, and a vow for

connected entities to improve healthcare facilities in the home, city, country and globe with the aid of a number of intelligent agents. SCH as a field of study at the intersection of public health, information system, big data, cloud computing, deep learning and artificial intelligence, has received a lot of attention from academia and industry.

U.S. National Science Foundation (NSF), as the most influential research and management organization in the world, has supported numerous scientific research projects that have led to global economic growth and the improvement of the quality of people's lives and health. Since 2012, NSF has supported the Smart Health and Wellbeing (SHB) Program. Later, it transfers to the Smart and Connected Health (SCH) program. As specified by SCH program solicitation [1], SCH program aims to "develop next generation healthcare solutions and encourage existing and new research communities to focus on breakthrough ideas in a variety of areas of value to health, such as sensor technology, networking, information and machine learning technology, decision support systems, modeling of behavioral and cognitive processes, as well as system and process modeling." There are 100 SCH projects that have been funded by NSF. Compared to scientific publications, the funded projects provide much more valuable information, which contribute to the hot themes and research challenges.

The purpose of this study is to adopt text analysis and text mining methods to analyze SCH projects funded by NSF, including what have been funded, characteristics of funded projects, and health problems and research challenges addressed by these projects. This study helps SCH researchers to understand the scope and characteristics of current NSF funded SCH projects so they can better prepare their NSF proposals. It also provides a case study to data science students and educators on how text analysis can be conducted for specific purposes.

2 General Characteristics: A Descriptive Analysis

We collected data from NSF website using its advanced ward search page [4] with NSF program element code 8018 (the code for smart and connected health program) on March 31, 2018. As a result, we retrieved 146 records that include the metadata of 100 SCH projects funded by NSF from 2012 to 2017. One project may have more than one record, as NSF allows different institutions to file their proposals separately even they are collaborating on the same project. The 146 records were retrieved and downloaded into an Excel file. The 25 metadata of the records include: Award Number, Title, NSF Division, Program(s), Start Date, Last Amendment Date, Principal Investigator (PI), State, Organization, Award Instrument, Program Manager, End Date, Awarded Amount to Date, Co-principal Investigators (Co-PI), PI email address, Organization Information (street, city, state, zip code, phone), NSF Directorate, Program Element Code(s), Program Reference Code(s), ARRA Amount, and Abstract. For the purpose of our study, we conducted analysis on 10 of the metadata elements of the records. Table 1 is a sample NSF project record with the 10 elements. Below we report our general descriptive analysis of the 146 records.

Table 1. Selected Metadata Information of a Sample NSF-Funded Project

Project Element	**Example**
Title	EAGER: Synthesizing Notes from Electronic Health Records to Make Them Actionable for Heart Failure Patients
Program(s)	Smart and Connected Health
Start Date	06/15/2017
PI	Jodi Forlizzi
Co-PI	Carolyn Rose, John Zimmerman
Organization	Carnegie-Mellon University
Awarded amount	$316,000.00
State	PA
End Date	05/31/2019
Abstract	This Early-concept Grant for Exploratory Research aims to help patients and caregivers have increased access to electronic health information. The research focuses on the Electronic Health Record (EHR). …. The research investigates new and more effective presentations of this information to patients (e.g., graphic, abstracted, actionable). …

2.1 Number of Funded Projects over Years

We found that the number of SCH funded records is increasing over years, from 4 records in 2012 to 48 records in 2017, as indicated in the dotted line showed in Fig. 1. The only exception is 2016 when there were 3 fewer records than year 2015, but still more records than year 2014. In 2017, the number of SCH records achieved 48, a growth of 65.52% over 2012. Note this calculation does not reflect the actual growth rate of the number of projects, because some projects have multiple records due to simultaneous filing of NSF proposals. As indicated by 2018 solicitation, 8-16 projects can be funded per year in the future [1].

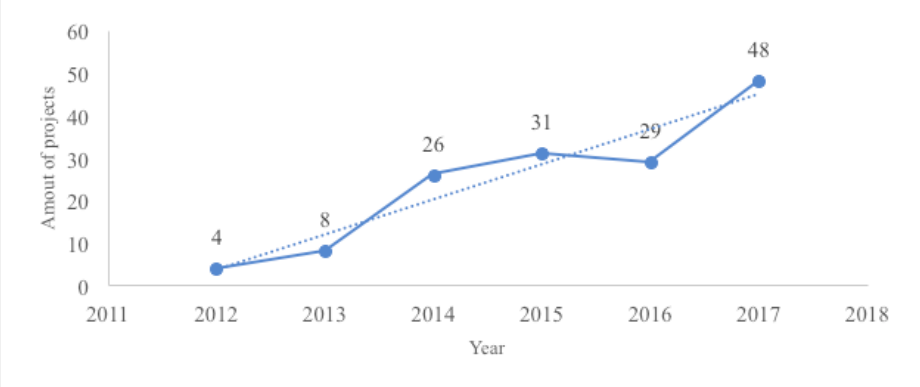

Fig. 1. Number of Funded Project over Years

2.2 Geographical distribution

The 146 records indicated that funded SCH projects were distributed in 35 states of the U.S. The top five states were Massachusetts, Texas, Pennsylvania, California and New York, all of which had more than 10 projects. Additionally, there were 4 states which had more than 5 projects, including Florida, North Carolina, Virginia and Maryland. Other states just had 1 or 2 projects. Some states, such as New Mexico and Hawaii did not have any project funded by NSF, however, that does not mean there were no researchers in these states involved in NSF funded SCH projects. Because NSF records only list PIs' states, it is very possible that some researchers participated as CO-PIs or major staff in SCH projects in those states are not listed.

2.3 Number of PI and CO-PIs

It is important to analyze the number of PI and CO-PIs to understand the situation of collaboration in SCH projects. The nature of SCH project demands that multidisciplinary teams work together to address multi-dimensional challenges ranging from fundamental science to clinical practice [1]. The distribution of the number of PI and CO-PIs were reported in Fig. 2.

NSF projects have only one PI, but can have multiple Co-PIs. In this study, we found that 51% of SCH records had two or more investigators. Among them, 34 records contain one PI and one CO-PI, 18 records having 2 CO-PIs, and 15 records having 3 CO-PIs. Furthermore, 5 records have 5 investigators (one PI and 4 CO-PIs) and 2 records have 6 investigators (one PI and 5 CO-PIs). The single PI projects may need more exploration. They may be part of a collaborative projects but filed the proposal separately, or maybe the PIs have a multidisciplinary research teams that have the required capabilities for conducting SCH projects.

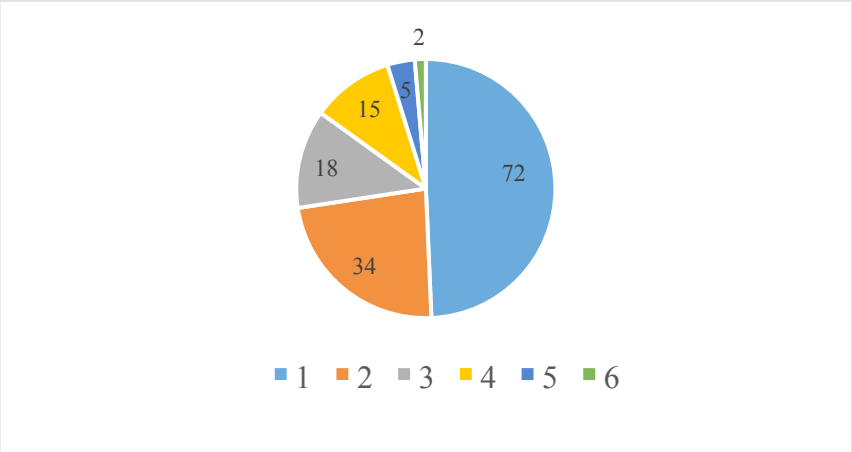

Fig. 2. Number of Investigators

2.4 Amount of Funds

For amount of funds, 15% of the 146 records are more than $1,000,000, 30% between $500,001- $1,000,000, 46% between $100,001- $500,000, 2% between $50,001- $100,000, and 7% less than or equal to $50,000. The amount of funds is showed in Table 2.

Table 2. Amount of Funds for Each SCH Record

Award Amount	Number of Projects	Percent (%)
More than $1,000,000	22	15
Between $500,001- $1,000,000	44	30
Between $100,001- $500,000	67	46
Between $50,001- $100,000	3	2
Less than or equal $50,000	10	7

In addition, we found that the amount of funds increases year by year, except for 2015 and 2017. In 2015, the total award amount was $14,474,172, which is $2,719,603 less than that of 2014. And the total award amount of 2017 was $18,980,644, which is 1,525,830 less than that of 2016. However, the numbers of SCH records in 2015 and 2016 are more than the previous year respectively, which indicates that the fund for each project on average was reduced in these two years. For 2018, the anticipated Funding Amount will be $11,000,000 to $20,000,000 [1].

2.5 Other Features

Organizations. There were 105 organizations that received at least one SCH award, or have at least one project funded by NSF on SCH. Among them, 22 organizations received 2 awards, 5 organizations received 3 awards, and 3 organizations received 4 awards. Our analysis indicates that 3 universities: Carnegie-Mellon University, Indiana University and University of Florida, have cultivated the most project teams in the study of SCH. Five other universities: Georgia Tech Research Corporation, Johns Hopkins University, North Carolina State University, University of Connecticut, and University of Virginia Main Campus, each of which had 3 funded SCH projects.

Duration of the Projects. The duration of majority projects (99 projects, 68%) is 3 or 4 years. Specifically, 63 projects were proposed to be completed within 4 years and 36 within 3 years. Further, 17 projects were proposed to be completed in 2 years and 15 projects should take 5 years to finish. There are only one 1-year project, and 5 projects to be completed in more than 5 years.

3 Term Frequency and Topic Analysis

3.1 Term Frequency and Word Cloud

The records we downloaded about NSF SCH projects includes titles and abstracts. They are well developed by the investigators containing most valuable information regarding the purposes, methods, diseases/health problems, and activities of the projects. After a review of the 146 records, we found that there are actually only 100 projects. Thirty-two projects have multiple entries (2 to 7) in the downloaded file. As explained earlier, NSF allows different organizations to file their proposals separately even they work on the sample project. Therefore, for the content analysis in this section and after, we removed the duplicate titles and abstracts in the download project file, leaving 100 titles and 100 abstracts for analysis.

We conducted automatic term frequency analysis using NLTK, the natural language processing toolkit [5]. After automatic processing and manual review of the most frequent 500 words, we obtained a list of words with their term frequency. The top 50 words are listed in Table 3. Fig 3 is the word cloud of the top 403 content words. The word cloud was created using wordart.com – a free word cloud generator [6].

Table 3. Most Frequent Terms in Titles and Abstracts

Term	Frequency	Term	Frequency	Term	Frequency
develop	257	learning	97	individual	64
health	225	disease	95	algorithms	61
patient	220	provide	88	behavior	61
model	211	approach	86	advance	60
system	204	improve	86	novel	59
technical	144	healthcare	74	sensor	59
clinical	122	support	71	potential	58
student	119	care	70	collaborative	57
medical	117	integrate	70	human	57
monitoring	103	information	66	real-time	57

We can make sense of the projects by observing the term frequency table and the word cloud: Most of the SCH projects are developing something, whether that is new device, new models, new technologies, or new processes; funded projects are well aligned with NSF program solicitation that focus on patient, health, medicine, student and care; many projects involve develop and use of systems and models. To make an accurate term frequency table, it is important to conduct stemming, or to normalize the different forms of words. For example, different forms for “model” can be “modeling,” “Modeling,” “models\,” and “Models” in the original abstract.

Fig. 3. Word Cloud of SCH Projects

3.2 Topic Analysis with Latent Dirichlet Allocation (LDA) Model

Latent Dirichlet allocation (LDA) model [7] has often been used to automatically identify the latent semantic topics in unstructured collections of documents. Giving a list of text documents, LDA model can identify topics in each document by a cluster of semantically related words [8][9]. Since LDA model can represent the document in a topic space instead of a word space, it helps to deal with the synonymy and polysemy problem from the semantic perspective and at the same time reduce the dimensionality. LDA has therefore been used in many semantic analytical researches such as: identification and monitoring the disruptive technologies [10], generating of the patent development maps [11], classification and pattern identification in patents [12] and so on. Specifically, Leah G. Nichols [13] proposed a topic model approach to explore the interdisciplinary of the NSF funding portfolio based on the NSF award and proposal database, which can help the NSF employees to better assess and administrate the funding portfolio, and the researchers to avoid duplicating others' research projects.

We conducted LDA analysis of the abstracts of the 100 projects. The purpose was to exam whether LDA could bring new insights to our understanding of the projects. Table 4 lists the word-level topics as identified by LDA. We used the open-source topic modeling tool gensim [14] for LDA analysis. Our program was configured to output 20 top terms for 20 topics.

It appears Table 4 displays a different set of important terms from what we could obtain for term frequency analysis in Section 3.1. For examples, disease names and health problem related terms such as "diabetes", "osteoporosis", "cardiac", and "cancer" are present in the list under different topics. Furthermore, we conducted LDA on bi-grams and tri-grams. More content terms were identified.

Table 4. The 20 Topics Generated by LDA Analysis

Topic ID	Terms under Each Topic
1	sleep, family, cpr, physical, feedback, therapy, child, researchers, obesity, evaluate
2	knowledge, behavioral, natural, sleep, smart, environmental, ontology, integration, researchers, transportation
3	dynamics, cardiac, environmental, computer, stateoftheart, imaging, multiscale, schemes, dimension, analyzing
4	management, gestures, sensors, user, specific, pis, researchers, behavioral, feedback, capacitive
5	software, motor, computer, function, pd, inspire, dynamics, cardiac, challenges, simcardio
6	detection, automation, physiological, early, clinicians, ad, methodologies, images, critical, education
7	mobile, personalized, emerging, advanced, study, asthma, researchers, ii, critical, computer
8	cancer, sleep, adaptive, intervention, effective, strategies, screening, breast, national, dynamic
9	imaging, guidelines, cognitive, chronic, specific, efforts, ultrasound, objective, multiple, effective
10	smart, imaging, mobile, diabetes, devices, ai, sensors, personalized, management, tools
11	imaging, failure, cardiac, software, device, simcardio, surgical, driving, fibrillation, source
12	postoperative, prosthesis, management, energy, agitation, family, smart, intervention, life, behavioral
13	dyadic, conference, dynamics, wellbeing, forum, psychotherapy, behavioral, finegrained, indicators, power
14	colon, mobile, education, imaging, aims, smart, behavioral, device, knowledge, undergraduate
15	surgical, outcomes, connectomics, conference, services, natural, mobile, social, forecasting, university
16	conference, dental, osteoporosis, informatics, international, services, doctoral, biomedical, collected, elderly
17	goals, inspire, personalized, coaching, outcomes, physicians, alerts, smart, adolescent, significant
18	sepsis, imaging, diagnostic, outcomes, cognitive, tests, enable, knowledge, ultrasound, cartilage
19	mathematical, theory, intervention, social, emergency, devices, therapy, effective, smart, tools
20	children, mobility, impairments, driving, agitation, dynamics, adhd, management, dyadic, imaging

4 Research Challenges Addressed by the Projects

One of the purposes of this study is to identify major research challenges that have been addressed by these SCH projects. Specifically, we would like to understand what diseases or health problems these projects have been tackling, and what popular research problems NSF investigators have been working on.

Two of the authors conducted a content analysis focusing on coding the projects (mainly the abstracts) on research problem/challenge, method/algorithm, disease, data, device, and other outcomes. The results of the analysis cannot be reported in this paper in detail due to the restriction on paper length. The content analysis helped us to achieve our purposes.

4.1 Diseases/Medical Problems Addressed by the Projects

The content analysis discovered that about 36 types of diseases or health problems have been tackled by the investigators, including respiratory diseases, infection plus systemic manifestations of infection, environmental public health issues, dementia, obesity, sickle cell disease, diabetes, cognitive disorders, mental trauma, heart problems, genetic diseases like cancer, sepsis, healthy life related problems, adolescent health, Attention-Deficit/Hyperactivity Disorder (ADHD) in teenagers and young adults, life threatening events in neonates, major complications following surgery, retinopathy of prematurity, strokes, epilepsy, depression, amputation, cardiovascular, traumatic brain injury, perioperative services, hepatitis C, alzheimer's disease, knee osteoarthritis, cognitive fatigue, psychotherapy, and children with mobility impairments. This list looks quite extensive. However, still many diseases or health problems are not included in these projects.

4.2 Research Areas and Challenges

Our analysis indicates that researchers are working in the following areas in smart and connected health:

1) Create novel methods and tools for the analysis of large-scale Electronic Health Record (EHR) data and social medial data to help diseases diagnose accurately, to improve patient care and/or to reduce costs;
2) Develop new or integrated methods, models, frameworks, and systems to help treat, monitor, or understand some diseases such as Asthma, Type-II Diabetes Mellitus (T2DM), Infection, and Heart Failure;
3) Develop new devices, mostly wearable sensors for disease monitoring, environmental control, injuries prevention, and safety;
4) Promote education in data science, training, and communication.

Understandably, different projects are dealing with different health problems. Based on the results of our term frequency analysis, topic analysis, and content analysis, we believe the following are the major research challenges investigated by these projects:

Electronic Health Record (EHR) data processing. At least 8 projects work on frameworks, integrating solutions, models, data mining approaches, and machine learning approaches to process EHR. Data processing is one of the major challenges in current big data environment and future smart health [15].
System design or computation model building. At least 22 projects emphasized system design and model building as one of their major objectives. Researchers work on developing systems and models to collect data and conduct data analysis.
Personalized or patient-centered medicine. At least 12 projects explore personalized or patient-centered medicine.
Training and education. At least 6 projects focus on student support for attending international conferences, institute support on global healthcare education.
Privacy preserving. At least 3 projects concentrate on exploring privacy preserving in EHR or other medical data.

5 Summary and Future Research

This paper analyzes 100 NSF projects that were identified as under the smart and connected health program based on their information retrieved from NSF website. Descriptive statistical analysis, topic analysis and content analysis were performed to understand the characteristics and research challenges tackled by these projects. SCH is a very important research area with many challenging research problems. Researchers who are interested in conducting SCH research will need to have collaborative spirit and be able to work as part of a team. We believe there are many opportunities for researchers to seek funding in NSF and other agencies in the area of smart and connected health.

This study is the beginning of our endeavor on smart and connected health. Our future research will be on two topics: One is to explore sophisticated text analysis techniques for effective and efficient understanding and mining of texts. The other is to initiate our NSF proposal application by tacking one of the interesting smart and connected health challenges.

References

1. Md IP, Raymond YKL, Haluk D, Md. AKA (2017) Smart health: Big data enabled health paradigm within smart cities. Expert Systems With Applications 87: 370–383.
2. National Science Foundation. Smart and Connected Health (SCH): Connecting Data, People and Systems. https://www.nsf.gov/pubs/2018/nsf18541/nsf18541.htm.
3. Clancy, CM (2006) Getting to “smart” health care. Health Affairs 25(6):589–592.
4. National Science Foundation Awards advanced Search. https://www.nsf.gov/awardsearch/advancedSearch.jsp
5. Bird, Steven, Ewan K, Edward L. Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit. http://www.nltk.org/book/.
6. WordArt.com homepage. https://wordart.com/, last accessed 2018/4/15.
7. Blei, David M, Andrew YN, Michael IJ (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993-1022.

8. Rosen-Z, Michal, Thomas G, Mark S, Padhraic S (2004) The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press,pp 487-494.
9. Steyvers, Mark, Padhraic S, Michal RZ, Thomas G (2004) Probabilistic author-topic models for information discovery. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM pp. 306-315.
10. Momeni, Abdolreza, Katja R (2016) Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. Technological Forecasting and Social Change 104: 16-29.
11. Kim, Mujin, Youngjin P, Janghyeok Y (2016) Generating patent development maps for technology monitoring using semantic patent-topic analysis. Computers & Industrial Engineering 98:289-299.
12. Venugopalan, Subhashini, Varun R (2015) Topic based classification and pattern identification in patents. Technological Forecasting and Social Change 94:236-250.
13. Nichols, Leah G (2014) A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics 100:741-754.
14. Gensim. https://radimrehurek.com/gensim/, last accessed 2018/4/15.
15. Olshansky SJ, Carnes BA, Yang YC, Miller N, Anderson J, Beltran-Sanchez H, Ricanek K (2016) The Future of Smart Health. Computer 49: 14-21.