Establishing Baseline in the Status of E-health Research in Norway

Andrius Budrionis, Karianne Lind, Inger Marie Holm, Omid Saadatfard, Rune Pedersen
Norwegian Centre for E-health Research
University Hospital of North Norway
Tromsø, Norway
e-mail: Andrius.Budrionis@ehealthresearch.no

***Abstract*—E-health is a rapidly evolving research field. In Norway, it is governed by the National E-health Strategy, defining six focus areas to be addressed by both academics and industry. The strategy is relatively new and much research in the field has been performed before it was developed. Literature search and machine learning classification methods were used to map scientific publications into the focus areas of the National E-health Strategy. Results showed that all strategic areas were represented in scientific publications; one focus area, new ways to provide healthcare, attracted the most attention from research communities. This paper presents a method and baseline of e-health research activities in Norway alongside the initial results of applying this method to the existing body of literature.**

Keywords-e-health strategy; publications; review; machine learning; classification; natural language processing.

I. INTRODUCTION

Telemedicine and e-health are priority research areas in Norway. Established academic communities have a long history of research and development in the field with success stories of functional telemedicine services used in clinical practice as early as 1991 [1]. Focus on e-health was strengthened in 2016 by establishing a dedicated agency under the Ministry of Health and Care Services, The Norwegian Directorate of E-health, with the goal of establishing and managing standards and national e-health solutions that contribute to high quality and effective health services.

Norwegian government and Norwegian Directorate of E-health are planning a shared national Electronic Health Record (EHR) solution for the entire healthcare sector, which is in line with the recommendations made in the policy document One Citizen – One Health Record [2][3] published by the Ministry of Health and Care Services in Norway. This policy document outlines an important strategic direction for the healthcare sector and recommends close collaboration between a number of stakeholders in the Norwegian e-health field.

The National E-health Strategy [4] and action plan 2017-2022 [5] describes the proposed strategic direction for the goal of a digitized and integrated healthcare system that provides a simpler, better and more comprehensive experience for the service recipients. The strategy is formed around six focus areas: 1) digitization of work processes, 2) better continuity of care, 3) better use of health data, 4) new ways to provide healthcare, 5) common foundation for digital services and 6) national e-health management and increased implementation. The first four are considered functional areas with direct value for healthcare services. The last two are considered foundations that are required for the first four to be realized.

Presenting Norwegian e-health research, categorized into these focus areas is helpful for policy makers, such as the Norwegian Directorate of E-health, to promote research in higher priority areas and coordinate activities in the national e-health field. Further, e-health research institutes and e-health organizations can use this information to search for research partners and to build collaboration networks. Currently, such information is not available. The objective of this study is to address this need by creating a classification of e-health research in Norway based on the six focus areas of the National E-health Strategy. This paper presents a method and baseline measures, scoping the state of e-health research activities in Norway based on scientific publications.

The reminder of this paper is organized as follows. Section II provides a summary of methods used to produce results, which are presented in section III. Section IV discusses the key findings and limitations of this work, while section V concludes the paper.

II. METHOD

This project could be divided into three phases: data collection, preprocessing and analysis. The reminder of this section summarizes the methods used in every phase of the project.

A. Data collection

To collect data on production of scientific publications within e-health, three major research databases (Scopus, Web of Science (WoS) and PubMed) were queried. Publications dated 01.01.2007 - 01.06.2018 were included. Publication search was performed in two phases between May and December, 2017:

1. Phase 1 included keyword-based search (predefined list of keywords and relevant MeSH terms) and author-based search covering a list of well-known researchers in the field.
2. Phase 2 was based on an extended author-based search including authors from publications identified in Phase 1.

1) *Phase 1*

All three databases (Scopus, Web of Science (WoS) and PubMed) were queried in Phase 1. We searched in title, abstract and keywords (Scopus and WoS) or in title and abstract (PubMed) and coupled the queries with affiliation

 ISBN: 978-1-61208-688-0

"Norway". In the author-based searches, we searched for author name combined with affiliation "Norway".

Phase 1 combined three approaches:

- MeSH term-based search (MeSH-term for e-health, telehealth and mHealth is "telemedicine").
- Search based on predefined list of e-health terms, listed below.
- Search for publications authored by well-known researchers in the field:
    - Researchers from the Norwegian Centre for E-health Research (E-health Research)
    - Known researchers within the Norwegian e-health field (identified previously)

Lists of e-health terms, expected to be present in the majority of e-health publications, were put together by senior researchers at the E-health Research. These terms were also combined with more general terms e-health, ehealth, telemedicine, telehealth and telecare.

The following list of e-health terms was used in the keyword-based search (? denotes a single random character, * denotes multiple random characters):

General keywords: e-health, ehealth, telemedicine, telehealth, telecare

EHR-related keywords: Electronic* Health Care Record, electronic*health record, electronic* healthcare record, electronic* medical record, electronic* patient record, patient health record, decision support, health information system, information infrastructure, information security, integration, process support, regionali?ation, semantic interoperability, standardi?ation, terminology, usability, privacy, archetypes, user interface.

Health analytics keywords: analy*, health analy*, large dataset*, big data, predictive analy*, computational epidemiology, health intelligence, artificial intelligence, machine learning, natural language processing, text and data mining, statistical analy*.

M-health keywords: Mhealth, m-health, homecare, sensor system*, medical app*, health app*, app*, mobile, remote monitoring, medical device, usability, information security, privacy, health record, self-management, wearable*, sensor*, self-generated data, economic impact*, facilitator*, mental health, tracking, empowerment, guidelines.

2) *Phase 2*

After phase 1 and deduplication, the dataset consisted of approximately 2000 references. Duplicates were removed following a simplified version of a method described by Bramer et al 2016 [6].

A list of the first and second author was compiled from the publications identified in Phase 1. Names, which were not searched for in Phase 1, were used to query Scopus database.

After de-duplication of the entire dataset it contained 3028 references. References were exported to a comma separated CSV file, including author, title, year, abstract, keywords, URL, DOI, Reference Type and Author Address.

B. Data preprocessing

Data analysis was performed in Python 3 environment using the latest versions of Natural Language Processing Toolkit (NLTK) [7] for free-text processing and scikit-learn [8] for analysis. Data analysis focused solely on publication title, keywords and abstract fields. Preprocessing included removal of stop-words and numeric values. Words in the free-text fields were stemmed (Snowball stemmer) and processed using Term Frequency – Inverse Document Frequency (TF-IDF) vectorizer. It resulted in a numerical representation of importance of n-grams (1 or 2 words in length) in the corpus.

A random sample of publications (N = 1700) was manually labelled assigning them to one of the 6 classes originating from the Norwegian National E-health Strategy [4]. References unrelated to e-health were also marked. Manual labelling was performed by one of 5 independent reviewers, who discussed classification criteria beforehand. Publications, which could not be classified by a single reviewer due to uncertainty regarding the correct class were discussed in common meetings where consensus class was determined.

Description of the classes in the strategy was used as classification criteria. Typical projects, which fit these classes are:

1) Digitization of work processes – improvement of work processes for healthcare professionals.
2) Better continuity of care – improvement of healthcare services for patients.
3) Better use of health data – health data analytics driven projects.
4) New ways to provide healthcare – novel services in healthcare, which were not available before.
5) Common foundation for digital services – infrastructure for large scale digital services.
6) National e-health management and increased implementation – national e-health solutions.

Labelled data were randomly split into training (80%) and testing (20%) data for supervised machine learning analysis.

C. Data analysis

Unsupervised machine learning methods were applied to cluster the publications and explore the data. In depth analysis was performed using supervised machine learning algorithms. Classification of publications was performed in two steps: binary classification (e-health/not e-health) and multi-class classification of e-health publications.

III. RESULTS

Results from data collection and analysis are presented in the reminder of this section.

A. Data collection

Data collection was performed in two phases including removal of duplicates. It resulted in 3028 publications, which were included in data analysis.

B. Supervised 2-class model

To clean the dataset from irrelevant publications, four binary classifiers (Linear SVC, Naïve Bayes, Logistic Regression and K-nearest neighbor) were trained and tested using 10-fold cross-validation on the labelled data (e-health/not e-health). Logistic regression classifier demonstrated the best performance for this problem and was

 ISBN: 978-1-61208-688-0

selected for further analyses. The performance of the 2-class classifier is presented in Table 1.

TABLE I. PERFORMANCE OF THE 2-CLASS CLASSIFIER

	Precision	Recall	f1-score
Not e-health	0.93	0.58	0.71
E-health	0.72	0.96	0.82

While interpreting the results, attention was directed towards the e-health class. The performance measures listed in Table 1 indicate, that the classifier is able to identify 96% of the relevant publications in the dataset (recall = 0.96), however, it also has a relatively high percentage of false positives in the e-health class (precision = 0.72). Regardless of the false positives rate, the classifier identifies almost all relevant e-health publications required for further analysis. The trained classifier was used to classify the unlabeled part of the dataset.

C. *Unsupervised model*

Data labelling stage could be perceived as biased due to overlap between the classes and human factors involved in the process. To adjust for the potential biases, unsupervised clustering of e-health publications was performed to check whether publication could be automatically assigned to a class characterized by content similarity. Principal Component Analysis (PCA) and k-means algorithm were fitted to visualize the high dimensional data in 3 dimensions (Figure 1).

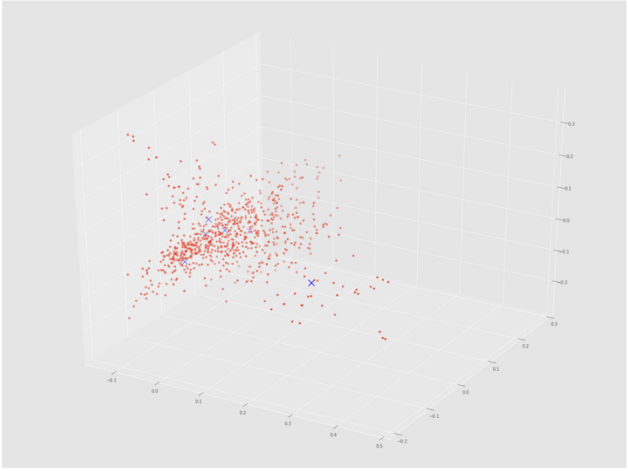

Figure 1. Clustering of e-health publications

Cluster analysis (before and after removing not e-health publications) identified no clear boundaries between the clusters and was not pursued further.

D. *Supervised 6-class model*

The 2-class model cleaned the dataset from the most of irrelevant publications (N = 1377). The cleaned dataset (N = 1651) showed uneven distribution of publication among classes with one class being overrepresented. The overrepresented class was down sampled in the training data to ensure that it is not overrepresented in the classification model. Four classifiers (Linear SVC, Naïve Bayes, Logistic Regression and K-nearest neighbor) were fitted and tested using 10-fold cross-validation, Naive Bayes model showed the best performance (Table 2). The trained classifier was used to classify the unlabeled e-health publications. Results are represented in Figure 2.

TABLE II. PERFORMANCE OF THE 6-CLASS CLASSIFIER

	Precision	Recall	f1-score
1. Digitization of work processes	0.70	0.58	0.63
2. Better continuity of care	0.61	0.62	0.62
3. Better use of health data	0.62	0.71	0.67
4. New ways to provide healthcare	0.74	0.77	0.75
5. Common foundation for digital services	0.53	0.62	0.57
6. National e-health management and increased implementation	0.66	0.64	0.65

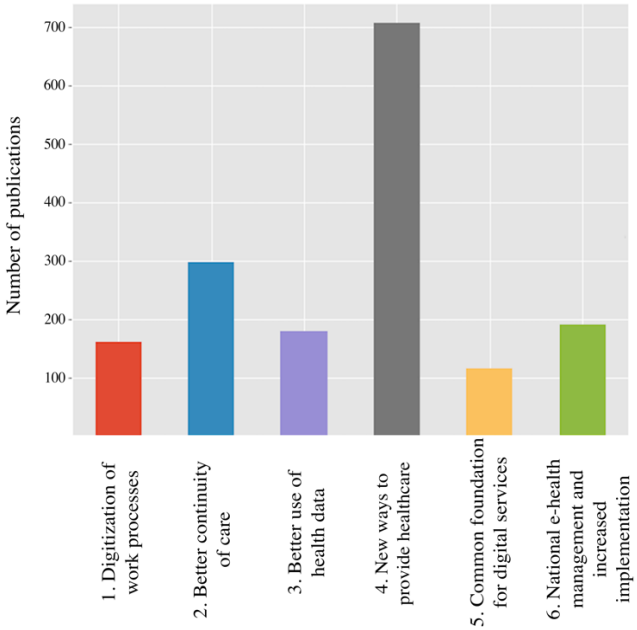

Figure 2. E-health publications classified according to the National E-health Strategy [4]

 ISBN: 978-1-61208-688-0

Top-10 keywords representing each class (sorted by importance):

1) Digitization of work processes: nurse, patient, use, care, hospital, electronic, record, health, work, support
2) Better continuity of care: patient, care, inform, health, communication, design, nurse, user, service, use
3) Better use of health data: data, patient, health, record, predict, inform, use, electronic, fall, medical
4) New ways to provide healthcare: diabetes, patient, health, use, social, service, care, technology, group, design
5) Common foundation for digital services: secure, health, standard, information, develop, service, data, ehealth, use, medical
6) National e-health management and increased implementation: telemedicine, health, information, implement, studies, infrastructure, technology, use, e-health, care

IV. DISCUSSION

Implications of the findings, methodological limitations and future works are discussed in this section.

A. Interpretation of findings

This project aimed to shed light on the status of e-health research activities in Norway with regards to the strategic documents [4][5]. The results of this project provide an overview on how the academical effort to bring the e-health field forward is reflected in scientific publications, and how these publications map to the focus areas defined in the National E-health Strategy (Figure 2). Disregarding the potential overlap between the focus areas and other methodological limitations, the project demonstrates the high pace of producing scientific results in the field. The most of research effort falls into class 4 (new ways to provide healthcare). This may not be a surprising finding, since the most of activities in e-health could be categorized as new ways to deliver healthcare.

The other publications distributed more evenly into the 5 classes. Class 5 (common foundation for digital services) had the weakest representation in scientific papers. This may be explained by the specifics of the Norwegian healthcare system. Norwegian healthcare is publicly owned and funded. Class 5 focus on the publications dealing with the infrastructure for delivering healthcare services to the citizens. Such infrastructure is partly developed from off-the-shelf components; procurement procedures are often based on other aspects than research (for instance, cost, reliability, flexibility, etc.). Space for research in this focus area is limited.

B. Methodological limitations

Methods to achieve the aforementioned results could be questioned. Level of uncertainty varied throughout the project, therefore, results should be interpreted in the context of the following limitations.

Data collection process was not strictly structured, therefore some publications may have been left out from further analyses. Phase 1 was focused on predefined keywords, which may not be completely representative for the entire field. Databases used in the search process do not include national publication channels, which often publish results in Norwegian.

Phase 2 used a list of the 1st and 2nd authors compiled from the publications identified in the Phase 1. All other authors were ignored. The search was performed only in Scopus database; PubMed and WoS added very few results in the previous phase. Data labelling process inherited the uncertainty originating from the National E-health Strategy. Focus areas are not defined to form mutually exclusive classes and there is a clear overlap between some of them. At the same time, publications often cover several focus areas and are difficult to assign to a single class. It was reflected in data labelling process, together with the additional uncertainty caused by human factors and background of the labelers. During the labelling process a class of e-health publications, which did not fit well into any focus areas was identified. Such publications deal with medical education, social media use for health purposes, reviews of various e-health topics and user health data storage solutions. These topics should be better addressed in the next versions of the e-health strategy.

Data analysis had to deal with the uncertainty inherited from the previous steps in the process. It may be reflected in relatively low precision and recall measures, especially in the 6-class model (Table 2). It may also be influenced by the false positives in 2-class model (precision = 0.72), which left some noise in the data filtering process. All the aforementioned aspects need to be taken into consideration when interpreting the results from this study.

C. Alternative classification strategies

The project started out with an idea to classify the publications according to the four research arenas at E-health Research: citizen services, patient pathways, health data and services for health professionals. The purpose was to establish a clear link between e-health research production in Norway and focus areas at E-health Research, which are logically distinct, and thus make classification easier. While the focus areas at E-health Research are well-aligned with the National E-health Strategy, they do not directly map to one another.

A decision to classify the identified publications according to the focus areas in the National E-health Strategy was a consensus reached in the project team. Regardless of the importance of measuring e-health research status towards national guidelines, focus areas defined in the strategy may not the best choice for classifying scientific production in the field. The National E-health Strategy is relatively new and much development in the field have taken place before it was made public. Focus areas are not meant to form mutually exclusive classes; research publications are often interdisciplinary and cover several focus areas. This situation causes uncertainty in classification, which might be avoided by selecting more distinctive classes. However, connection to the existing e-health strategy would be lost. One question that could be asked is whether the National E-health Strategy should be used to classify and potentially influence e-health research, or it is e-health research that should influence the e-health strategy?

 ISBN: 978-1-61208-688-0

D. Future work

This paper aimed to establish a baseline in the status of e-health research in Norway. Studying the development of the field is meant to be a continuous process, which is repeated periodically, preferably every third year. Results can be used as an input for the policy-makers in organizing, coordinating and allocating research funding, strengthening the weak focus areas and research institutions. Lessons learned during this iteration will contribute to a more structured and easier reproducible data collection and analysis process in future iterations. Further analyses will focus on classifying the publications according to the research institutions, mapping the focus areas of the National E-health Strategy to the interests of the most important academic actors in the country. The project aims to communicate the findings in a visual and easily understandable manner, therefore results will be represented as an interactive periodically updated map.

V. CONCLUSION

This paper presented a method and initial results scoping the status of e-health research in Norway based on scientific publications from the last decade. It mapped the published research results to the focus areas of the National E-health Strategy. Findings show that all focus areas are represented in the previous and ongoing research activities. Most of the publications fell into the focus area dealing with "new ways to provide healthcare services". The focus area "common foundation for digital services" had the weakest representation in the identified scientific publications.

REFERENCES

[1] I. Nordrum et al., "Remote frozen section service: A telepathology project in northern Norway," *Human Pathology*, vol. 22, no. 6, pp. 514–518, Jun. 1991.

[2] Norwegian Ministry of Health and Care Services, "Meld. St. 9 (2012–2013)," *Regjeringen.no*, 30-Nov-2012. [Online]. Available: http://www.webcitation.org/73nsLIeoU. [Accessed: 09-Nov-2018].

[3] Norwegian Directorate of E-health, "Utredning av «Én innbygger – én journal»," [Report "One Citizen - One Health Record"] Dec-2015. [Online]. Available: http://www.webcitation.org/73nstSwCb. [Accessed: 09-Nov-2018].

[4] Norwegian Directorate of E-health "Nasjonal e-helsestrategi 2017-2022." [National e-health strategy 2017-2022] [Online]. Available: http://www.webcitation.org/73ntEa1LR. [Accessed: 09-Nov-2018].

[5] Norwegian Directorate of E-health, "Nasjonal handlingsplan for e-helse 2017-2022." [National action plan for e-health 2017-2022] [Online]. Available: http://www.webcitation.org/73ntLCceo. [Accessed: 09-Nov-2018].

[6] W. M. Bramer, D. Giustini, G. B. de Jonge, L. Holland, and T. Bekhuis, "De-duplication of database search results for systematic reviews in EndNote," *J Med Libr Assoc*, vol. 104, no. 3, pp. 240–243, Jul. 2016.

[7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

 ISBN: 978-1-61208-688-0