23.11.2022

Data-driven identification of elderly individuals with future need for multi-sectoral services (MAITE-STUDY)

- research plan

Finnish Institute of Health and Welfare (THL)
Juha Koivisto, Principal Investigator
Teemu Keski-Kuha, Senior Specialist

VTT Technical Research Centre of Finland
Emmi Antikainen, Research Scientist
Jaakko Lähteenmäki, Principal Scientist
Juha Pajula, Research Scientist
Heba Sourkatti, Research Scientist
Mika Hilvo, Research Team Leader

Päijät-Häme Joint Authority for Health and Wellbeing (PHHYKY)
Corinne Soini, Division Director Customer Guidance
Marika Jalonen, Chief Data Officer

Revision history

Date	Author	Description
14.2.2022	Jaakko Lähteenmäki	Version annexed to the submitted research permit application.
21.11.2022	Jaakko Lähteenmäki	Updated version, main changes: - death year added to Table 1 - customer group added to Table 1 - RAI indicator list updated in Table 1 - research group updated with new members

Contents

1 Introduction

The aging population causes a growing need for health and social services and the related costs are increasing rapidly. At the same time there is an increasing shortage of professional workforce. Based on earlier research, 10% of the Finnish population consume 80% of all costs incurred in healthcare and social services. Furthermore, a specific subgroup of this 10% consume an exceptionally large number of different health and social services [1]. These individuals with increased need for multi-sectoral services are often not recognized properly in health and social services settings although they would benefit from integrated care and coordinated interventions. Early identification of individuals with increased need of multi-sectoral services would enable proactive actions with potentially high positive impact on quality of life and costs of health and social services.

Artificial intelligence (AI) based prediction of health and social services usage has been proposed in several earlier studies. Typically, the underlying goal has been to provide information needed to improve efficiency and quality of healthcare processes. Most cases involve prediction of hospital readmissions and emergency department visits (e.g. [2]–[5]). Examples of other addressed areas include exacerbation of chronic disease patients (e.g. [6][7]), future diagnoses (e.g. [8]–[10]), and hospital no-show appointments (e.g. [11]).

While the majority of the existing studies focus on predicting health risks of patients, less attention has been given to the overall use of health and social services. Existing studies in this field focus on the prediction of healthcare costs. A study using Finnish Institute for Health and Welfare (THL) data resources (Avohilmo) presented a risk adjustment model using individual trajectories of disease diagnoses and treatments for individual-level prediction of future health and social services costs [12].

Only recently, there have been specific efforts towards developing data-driven prediction tools to support public healthcare and social service providers. Such a pilot project for predicting “undesired” endpoints for elderly citizens has been reported by the Kainuu Social and Health Care area[1]. Similar projects have been carried out also in Eksote[2] and Päijät-Häme[3] areas. From these pilot projects, only limited public information is available currently.

The goal of this study - building on the experiences of Kainuu, Eksote and Päijät-Häme – is to develop an AI-based model to identify elderly individuals with increased need for multi-sectoral services in the future. The challenge of such a model is that usage of services is in many cases inevitable. Actually, the individual’s adherence to the recommended care will likely lead to a better patient outcome with reduced need for services in the long term.

In order to address this challenge, we will need to take into account the diversity of the elderly population and their profiles as service consumers. Thus, we will not use the number of different services directly as an endpoint. Instead, group-specific endpoints, such as uncontrolled diabetes complications, will be defined. These “surrogate” endpoints are expected to correlate with the overall endpoint. We will specifically look for actionable endpoints, referring to undesired outcomes, which could be avoided by proactive actions. We will also analyse the correlation between the surrogate endpoints and the overall endpoint (usage of multiple different services).

This study is part of a joint research project (“Monialaisen palvelutarpeen ennakointi tekoälyn avulla – kansallinen kehittämispilotti ikäihmisten asiakasryhmää koskien”) coordinated by the National Institute of Health and Welfare (THL) and participated by VTT Technical Research Centre of Finland, Päijät-Häme Joint Authority for Health and Wellbeing (PHHYKY), University of Helsinki and University of Lappland. The main funder of the project is the Ministry of Social Affairs and Health (STM).

The study will be carried out by VTT, THL and PHHYKY. Data resources of PHHYKY will be used.

[1] https://digifinland.fi/wp-content/uploads/2019/11/Tietojohtamisen-pilotti-SoteDigi-Oy-Kainuun-sote-loppuraportti-pdf.pdf
[2] https://www.eksote.fi/eksote/ajankohtaista-ja-mediatiedotteet/2021/Sivut/Yhteinen-tulevaisuus-%E2%80%93-Eksoten-hyvinvointiblogi-Hyv%C3%A4n-palvelun-takana-on-laadukas-ja-ajantasainen-tieto-.aspx
[3] Ismo Rautiainen: Monialaisen tuen asiakkaan asiakasohjauksen kriittiset menestystekijät Päijät-Hämeen hyvinvointiyhtymässä

2 Objectives

We have identified two (qualitative) research hypotheses:

1. An elderly individual's future wellbeing status (endpoints) can be predicted based on data registered during the earlier use of health and social services

2. The elderly individual's undesirable wellbeing situation is associated with the need for multi-sectoral services (overall endpoint).

The overall aim of the study is to test these hypotheses and to develop and validate AI-based (machine learning) models based on them. Based on individual-level data in health and social data registers, the developed models support early identification of the elderly individuals with increased need for multi-sectoral services.

The detailed objectives of the study are:

- Identifying undesired outcomes (endpoints) relevant to specific subgroups of the elderly population
- Identifying appropriate data elements documented in the customer and patient information systems associated with the individual's actionable health and social problems
- Development of machine learning models using longitudinal data for identifying individuals with one or more undesired outcomes in the future
- Assessment of the performance of the developed machine learning models
- Assessment of the correlation between the endpoints and the increased need for multi-sectoral services

3 Research design and methods

3.1 Study design

This is an exploratory study based on the analysis of retrospective data for identifying elderly individuals with increased need for multi-sectoral services. The study employs machine learning methods using data in customer and patient information systems to classify individuals with respect to undesired outcomes (endpoints) which will lead to the use of multiple different services in the future (overall endpoint).

The inclusion criteria for the study are:

- Permanent residence in PHHYKY region during the time period 2018-2021
- Year of birth in range 1932-1952 (current age: 70-90)
- Used PHHYKY services at least once during the year 2018

The exclusion criteria for the study are:

- Death during the time period 2018-2021 (if known)
- Move out from PHHYKY service area during the time period 2018-2021 (if known)

The endpoints will be defined during the early phase of the study based on statistical analyses of the available customer and patient data and discussions with health and social services professionals. We will focus on actionable endpoints referring to undesired outcomes, which could be avoided by

proactive actions. Examples of possible endpoints are: (1) uncontrolled diabetes, (2) uncontrolled hypertension, (3) untreated substance use disorder, (4) untreated mental illness, (5) independent living not adequately supported, and (6) other health related problem not treated. The overall endpoint is the individuals' increased usage of multi-sectoral services. Exact, quantitative definition of the endpoints will be made in the early phase of the study in co-operation with healthcare and social service professionals of PHHYKY.

3.2 Data resources

Data resources of PHHYKY registers available via the PHHYKY data warehouse environment will be used. The resources to be used are:

- Patient and customer information systems data (e.g. originating from Effica and Lifecare)
- Service needs assessment data (e.g. originating from Raisoft)

The data contents to be used shall include only structured data. Data in free text form will not be used. The identified data contents planned to be used in the study are listed in Table 1. The availability of medication data has not yet been secured, as the data warehouse environment is still under development. The study would benefit from the listed medication data, but can be performed without them if necessary.Update (21.11.2022): medication data is available and will be used as specified in Table 1.

The identified data contents for the period of 1.1.2018 – 31.12.2021 will be used.

Table 1. Data contents to be used in the project.

Data resource	Data definition
Health and social services encounter data (Effica, Lifecare)	**Basic demographic information:** age, sex, municipality of residence, death year **Information about health and social services visits**: - customer group - primary and specialized healthcare outpatient and inpatient visits - mental health services - substance abuse services - social services for elderly - dental care services For each visit: • Visited service. Based on PTHAVO Palvelumuoto[1] classification code or similar. • Date of visit (+duration in case of inpatient visit) • Diagnoses documented for the visit (‘päädiagnoosi’) (ICD10 code). • Operation (procedure) documented for the visit (‘toimenpide’) (national operation code) • Medication (‘lääkitys’). Usage of the following drug groups: M01 (‘tulehduskipu- ja reumalääkkeet’), N05 (‘psykoosi- ja neuroosilääkkeet sekä unilääkkeet’), N06 (’masennuslääkkeet ja keskushermostoa stimuloivat lääkeaineet’) • Laboratory measurements: B -HbA1c (6128), P -Gluk (1471), B -Hb (1552), P -Ferrit (4826), S -D-25 (1220), B -PEth (12510)
Social services decisions (Effica, Lifecare)	**Information about social services decisions:** - rehabilitation - elderly housing service (‘asumispalvelut’) - caregiver support (‘omaishoidon tuki’) - homecare - transportation services For each decision: - service decision category (‘palvelupäätös’) - decision outcome, positive/negative - date
Service need and physical function assessment (‘palvelutarpeen ja toimintakyvyn arvioinnit) (Raisoft)	**Information about service need and physical function assessments (for any purpose):** RAI indices: Activities of Daily Living Hierarchy (ADLH) Instrumental Activities of Daily Living Summary Scale (IADLP21) Cognitive Performance Scale (CPS) Depression Rating Scale (DWS)* Positive Symptoms Scale (PSS) * Negative Symptoms Scale Short (NSS)* Changes in Health, End-stage disease and Symptoms and Signs (CHESS) Communication (COMM) Body Mass Index (BMI) Rehabilitation given by nurse; ‘Hoitajien antama aktiivinen kuntoutus’ (HAAKu) Method for Assigning Priority Levels; 15 classes (MAPLe15) Method for Assigning Priority Levels; 5 classes (MAPLe5)

* Index in use since 2019

3.3 Data Analysis methods

The study employs machine learning methods in two stages. First, a preliminary cluster analysis will be carried out to identify candidate endpoints for final selection with healthcare and social professionals. This analysis will also reveal the combination of variables to be used in converting the qualitative endpoints into quantified form. In the second phase, different machine learning techniques (such as transformer neural networks) will be applied for predicting the occurrence of the outcomes based on longitudinal health and social services data.

The prediction model will be validated using cross-validation methods. We will evaluate the association of the various surrogate endpoints with the individual usage of multiple different services. Standard statistical methods, such as logistic regression, will be used for this.

3.4 Sample size

Based on the PHHYKY area total population (about 205 000) and age distribution of the Finnish population (about 15% in age range 70-90) the target group size could be about 30000 individuals. Based on national statistics on yearly primary care customers[2], we estimate that at least 70% of the elderly population would have used primary healthcare services during one year (in this case year 2018). Consequently, we estimate that about 21 000 individuals would fulfil the inclusion criteria.

4 Schedule

The targeted study schedule is depicted below. Taking into account possible delays in the data acquisition and in results publication, the end-date to be used in the data permit application is 31.12.2024.

Preparatory activities include definition of the study endpoints in co-operation with health and social services professionals of PHHYKY. The preparatory phase also includes all necessary actions needed to fulfil the requirements for accessing the data in PHHYKY's healthcare and social data register. Data pre-processing phase includes the tasks of verifying the integrity of the study data and organizing it in the research database. The data analysis and AI-model development phase includes the activities of identifying specific risks groups and tuning related prediction models. Model validation phase includes evaluating the performance of the models based on cross-validation techniques. The final phase is devoted to drawing conclusions and publishing the study results. The target is to publish at least one scientific journal article. Additionally, study results will be presented in the MAITE-project workshops and other events to ensure dissemination to other stakeholders, in particular other wellbeing services counties.

Study phases	2022												2023					
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
Preparatory activities																		
Data pre-processing																		
Data analysis and AI-model development																		
Model validation																		
Results integration and publication																		

5 Limitations

The study has been limited into one wellbeing services county (PHHYKY) due to practical reasons - with the available project resources it would have not been possible to execute a larger study. Due to

[2] https://www.julkari.fi/bitstream/handle/10024/137690/SH%202018_11%20Suomalaisten%20sosiaali-%20ja%20terveyspalvelujen%20k%c3%a4ytt%c3%b6%20tilastojen%20valossa%20_%20L%20Kestil%c3%a4%20et%20al.pdf?sequence=1&isAllowed=y

this limitation the available sample size will be relatively small. The challenge will be that the foreseen study population of 21000 divides into a number of smaller subgroups with specific surrogate endpoints, which need to be separately analyzed. Furthermore, only a part of the individuals can be used for model training, while at least 25% need to be saved for validation and testing of the AI-model. We still believe that the available sample size allows development of an initial AI-model with sufficient predictive power to establish a basis for further development in subsequent studies powered by data from larger areas.

6 Ethical aspects

The study will be based entirely on existing register data. Patients will not be contacted in any phase of the study and patients will not be subject to any additional tests or investigations. The results of the study will not directly affect the therapy of the individual patients.

Data permit application will be made to PHHYKY as the register controller. In line with the Secondary use of health and social data act (552/2019), PHHYKY is entitled to grant the data permit for this study, which uses solely PHHYKY's own data resources.

Based on the guidance by Finnish National Board on Research Integrity (TENK guidelines[3]), the research setting falls to the category of register-based studies where ethical review statement from a human sciences ethics committee is not required. However, the TENK guidelines will be followed to ensure compliance with good ethics practices throughout the lifecycle of the study.

In particular, the following practices will be followed:

- **Minimization of data.** The data contents needed for the study are carefully analysed in co-operation with health and social services professionals before the data permit application and only data resources necessary for the study are requested.

- **Data extraction.** PHHYKY as the register controller will be responsible for extracting the data to be used in the study from their health and social data registers. PHHYKY will transfer the data to the secure processing environment (Kapseli[4]) provided by Findata as required by the legislation (552/2019). Standard Findata process will be used for the transfer. Before the data transfer, PHHYKY will ensure that the material does not contain direct person identifiers.

- **Data processing environment.** Kapseli environment maintained by Findata is a highly secure processing environment and does not allow data to be downloaded for local processing. Data will be processed in the Kapseli environment from VTT's workstations by using a remote desktop application.

- **Data protection practices at VTT.** The study will be carried out at VTT as a part of the MAITE project, conducted along VTT's standard project practices based on certified quality system (ISO 9001). Even though direct person identifiers are removed, the data are still treated as sensitive personal information as re-identification is theoretically possible. Specific security practices will be followed to ensure high security in processing the data. This includes, for example, ensuring that results to be published will not include information that could enable identification of study participants. Access to data will be enabled solely to the researchers of the project group listed in this research plan. All data will be removed after the MAITE project has been completed and the results have been published. Data will be removed latest by 31.12.2024.

7 Study group

The study group includes researchers of THL, VTT and PHHYKY:

[3] Ihmiseen kohdistuvan tutkimuksen eettiset periaatteet ja ihmistieteiden eettinen ennakkoarviointi Suomessa (tenk.fi)
[4] Kapseli remote access - Findata

Finnish Institute of Health and Welfare (THL)

Juha Koivisto, DSocSci, Docent, Principal Investigator

Teemu Keski-Kuha, M.Sc. (Tech), Senior Specialist

VTT Technical Research Centre of Finland

Emmi Antikainen, M.Sc. (Tech), Research Scientist (participated until 31.7.2022)

Jaakko Lähteenmäki, Lic.Sc. (Tech), Principal Scientist

Juha Pajula, PhD (Tech), Senior Research Scientist

Heba Sourkatti, MSc (Tech), Research Scientist

Mika Hilvo, PhD (Tech), Research Team Leader

PHHYKY

Corinne Soini, Division Director Customer Guidance

Marika Jalonen, Chief Data Officer

VTT's researchers will be responsible for carrying out the main part of the study, in particular the tasks of data analytics, model development and testing.

THL's researchers will, together with VTT researchers, contribute to the definition of the study plan and objectives as well as to the analysis of study results.

PHHYKY experts will support the study in defining the study endpoints, in selecting the data elements to be used and in preprocessing the data to a usable form for analysis. In these activities PHHYKY is supported by 2M-IT personnel.

All three parties will contribute to the analysis of the results and drawing the conclusions.

8 Costs and funding

The study will be carried out as part of the MAITE project.

The budgeted VTT's costs for performing the study are 180 000 €. The budget is covered by grant funding of STM (75%) and VTT (25%).

The budgeted THL's costs for performing the study are 9000 €. The budget is covered by grant funding of STM.

The budgeted PHHYKY's costs for performing the study are maximum of 15000 €. The budget is covered by grant funding of STM.

9 Significance of results

This exploratory study is expected to produce initial results complementing those achieved in the pilots mentioned above in section 1. The study aims to develop a predictive model for the identification of elderly individuals with an increased risk of becoming users of a large number of different health and social services. Such model could potentially be used in several ways.

In the case of customer encounters, the model would enable decision support for the healthcare or social services planning. Early identification of individual risks would enable preventive and corrective actions pertaining to the individual's care and services to be taken at early stage. On the other hand, applying the model for larger populations would enable knowledge-based management of health and social services and preventive interventions to be targeted to specific risk groups. In both cases, it is likely that savings in health and social services and improvement in quality of life could be achieved.

The importance of preventive and proactive services has already been widely recognized, but services are still mostly reactive rather than proactive and preventive. One challenge may have been the difficulty to access and exploit data resources. Currently this is changing, as the healthcare providers are investing in data lake systems, which enable more efficient and timely exploitation of data. Also, the large-scale social welfare and healthcare reform being currently implemented in Finland is seen as an opportunity to improve service quality and efficiency. The welfare service counties are expected to have more resources to invest in proactive and preventive services models, from which the benefits will be visible after several years. Additionally, substantial national effort has been focused for harmonized tools and practices enabling data-driven and knowledge-based management[5].

The MAITE study is in a good position to produce useful results to be later taken to operational tools in healthcare and social welfare services. Applicability of the study results is expected to be further improved thanks to additional parallel activities carried out in the MAITE project. The project addresses the challenges coming from legislation, which restricts the use of individual-level data for proactive interventions where the individual's data is processed and the individual is contacted without an existing service episode. The MAITE project also considers the aspects related to information system architectures needed to exploit data-driven models in the context of health and social services.

[5] https://stm.fi/en/project?tunnus=STM029:00/2020

10 References

[1] J. Koivisto and H. (toim. . Tiirinki, “Monialaisen palvelutarpeen tunnistaminen sosiaali-, terveys- ja työvoimapalveluissa,” Aug. 2020, Accessed: Dec. 23, 2021. [Online]. Available: https://julkaisut.valtioneuvosto.fi/handle/10024/162382.

[2] A. C. M. Amato, R. V. dos Santos, D. Z. Saucedo, and S. J. de T. A. Amato, “Machine learning in prediction of individual patient readmissions forelective carotid endarterectomy, aortofemoral bypass/aortic aneurysm repair, andfemoral-distal arterial bypass,” *SAGE Open Med.*, vol. 8, p. 205031212090905, Jan. 2020, doi: 10.1177/2050312120909057.

[3] W. Liu *et al.*, “Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding,” *PLoS One*, vol. 15, no. 4, Apr. 2020, doi: 10.1371/JOURNAL.PONE.0221606.

[4] Q. Xie *et al.*, “Effect of ABCB1 Genotypes on the Pharmacokinetics and Clinical Outcomes of New Oral Anticoagulants: A Systematic Review and Meta-analysis,” *Curr. Pharm. Des.*, vol. 24, no. 30, pp. 3558–3565, Oct. 2018, doi: 10.2174/1381612824666181018153641.

[5] M. Loreto, T. Lisboa, and V. P. Moreira, “Early prediction of ICU readmissions using classification algorithms,” *Comput. Biol. Med.*, vol. 118, p. 103636, Mar. 2020, doi: 10.1016/J.COMPBIOMED.2020.103636.

[6] G. Luo, B. L. Stone, X. Sheng, S. He, C. Koebnick, and F. L. Nkoy, “Using Computational Methods to Improve Integrated Disease Management for Asthma and Chronic Obstructive Pulmonary Disease: Protocol for a Secondary Analysis,” *JMIR Res. Protoc.*, vol. 10, no. 5, May 2021, doi: 10.2196/27065.

[7] D. W. Mapel *et al.*, “Development and Validation of a Healthcare Utilization-Based Algorithm to Identify Acute Exacerbations of Chronic Obstructive Pulmonary Disease,” *Int. J. Chron. Obstruct. Pulmon. Dis.*, vol. 16, pp. 1687–1698, 2021, doi: 10.2147/COPD.S302241.

[8] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, “Learning to Diagnose with LSTM Recurrent Neural Networks,” *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, Nov. 2015, Accessed: Dec. 23, 2021. [Online]. Available: https://arxiv.org/abs/1511.03677v7.

[9] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” *NPJ Digit. Med.*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/S41746-021-00455-Y.

[10] Y. Li *et al.*, “BEHRT: Transformer for Electronic Health Records,” *Sci. Reports 2020 101*, vol. 10, no. 1, pp. 1–12, Apr. 2020, doi: 10.1038/s41598-020-62922-y.

[11] S. AlMuhaideb, O. Alswailem, N. Alsubaie, I. Ferwana, and A. Alnajem, “Prediction of hospital no-show appointments through artificial intelligence algorithms,” *Ann. Saudi Med.*, vol. 39, no. 6, p. 373, 2019, doi: 10.5144/0256-4947.2019.373.

[12] Y. Kumar *et al.*, “Predicting utilization of healthcare services from individual disease trajectories using RNNs with multi-headed attention,” *Machine Learning Research*, vol. 116. PMLR, pp. 93–111, Apr. 30, 2020, Accessed: Dec. 23, 2021. [Online]. Available: https://proceedings.mlr.press/v116/kumar20a.html.