Computing and Informatics, Vol. 41, 2022, 233–252, doi: 10.31577/cai_2022_1_233

HOW TO OVERCOME LACK OF HEALTH RECORD DATA AND PRIVACY OBSTACLES IN INITIAL PHASES OF MEDICAL DATA ANALYSIS PROJECTS

Yehya Mohamad

Fraunhofer Institute for Applied Information Technology FIT
Schloss Birlinghoven
53757 Sankt Augustin, Germany
e-mail: yehya.mohamad@fit.fraunhofer.de

Alexander Gabber, Sonja Heidenblut

Lehrstuhl für Rehabilitationswissenschaftliche Gerontologie
Humanwissenschaftliche Fakultät
Universität zu Köln, Herbert-Lewin-Strasse 2
50931 Köln, Germany
e-mail: agabber@uni-koeln.de, sonja.heidenblut@uni-koeln.de

Daniel Zenz

Smart-Q GmbH, Lise-Meitner-Allee 4
44801 Bochum, Germany
e-mail: zenz@smart-q.de

Anam Siddiqi, Henrike Gappa

Fraunhofer Institute for Applied Information Technology FIT
Schloss Birlinghoven
53757 Sankt Augustin, Germany
e-mail: {anam.siddiqi, henrike.gappa}@fit.fraunhofer.de

Abstract. The lack of electronic health record data in general and especially at initial phases of medical research projects is common and is one of the main reasons for delay or failure of such projects. One of the health areas with little attention is the home care area, where patients are being supported by their families or informal caregiver at home. In this paper we present related work on medical data formats and synthetical data generation of medical health records. Furthermore, it presents an approach to generate synthetic electronic health records (HER) that are readily available; suited to research; and free of legal, privacy, security and intellectual property restrictions to be used in home care research projects. We adapted and used Synthea™, an open-source software framework that simulates the lifespans of synthetic patients to generate synthetic EHRs. This paper presents the use case of home care from the capturing of user requirements of home care patients, translating the requirements into a data model, feeding the data model into Synthea™ framework, which produces synthetical health data records mainly as QuestionnaireResponse instance of the Fast Healthcare Interoperability Resources (FHIR) to using these EHRs to build an initial machine learning data model for home care.

Keywords: Health record, FHIR, HL7, home care, machine learning, Questionnaire, QuestionnaireResponse, Synthea™

1 INTRODUCTION

Access to high quality synthetical health datasets is essential for medical research projects in areas that are not directly oriented to clinical treatments, with no clinical or medical implications, including software development, testing and clinical training. Such large repositories are required for testing of interoperability, data analytics or machine learning modelling.

Legal, privacy, security and intellectual property restrictions prevent developers in research projects from access to real EHRs. Where access to real datasets is not possible for the abovementioned reasons or simply for the lack of such data, developers have usually used anonymized EHRs. Anonymized EHRs are meanwhile commercially available from different providers e.g. clinics. The use of anonymized real EHRs is accompanied with issues of privacy, confidentiality and consent. The amount, quality and suitability of anonymized EHRs for different use cases is by far not sufficient due to high risk of harm from public disclosure and identification of individuals resulting from the release leading to re-identification of these records. As a result, issues regarding legal concerns and policy frameworks effectively prevent progress of research projects in eHealth. To overcome these obstacles, an international research collaboration has developed an approach, method and open-source system for generating realistic synthetic EHRs [5]. Based on this work a numerous of further projects have used the open-source results to generate synthetic EHR data and to use these in different use cases such as home care use case described in this paper.

In our nationally funded research project INGE integrate4care (digitale INtegrierte GEsundheits- und Pflegeversorgung mit IT-gestütztem Pflegeberatungsbesuch nach § 37.3 SGB XI/Digital integrated health and home care with IT-supported in-home care consultancy in conformance to § 37.3 social security statute book XI) we faced the lack of data problem as described above. It was one of the goals in this project to implement an early warning system based on a risk assessment of people who were taken care of by informal caregivers at home. The early warning system was focused on predicting the risk of deteriorating health conditions of care recipients, e.g., declining cognitive skills and the burden on informal caregivers. Machine learning algorithms were foreseen to apply results from analysis of home care histories to an individual home care setting and support the in-house care consultancy by predictive warnings and, if possible, measures for prevention. For this purpose, a suitable repository with histories of home care data would have been necessary, however, since it was known that such repositories do not exist yet, it was planned that several mobile nursing services will collect such data in course of the project with a specifically developed app (see Section 3 for further details). However, due to the COVID-19 pandemic in-home consultancy took place significantly less for security reasons. To overcome the aforementioned constraints synthetic data generation was used to compensate for the lack of care data.

2 RELATED WORK

2.1 Medical Data Formats

Currently the Fast Healthcare Interoperability Resources (FHIR) (Release 4 (1st Normative Content + Trial Use Developments)) [17] is the main medical interoperability standard, that is being maintained by Health Level 7. Grahame Grieve [12] started the development in 2010 supported later by HL7, an experienced developing group of standards for the healthcare sector. The standard is open source [6], which provides on the one side developers a good opportunity to use freely the standard and on the other side the advantage for HL7 to get early feedback from developers during the actual development stage and the standard can therefore be improved incrementally. For the industry most standards are not open source, associated with high license costs and therefore, changes of utilized standards used to be a hard decision and the change itself very difficult. The FHIR standard is targeting a wide range of healthcare institutions, from clinical and public health laboratories over healthcare institutions, like hospitals or home care, to vendors, e.g., clinical decision support systems vendors, or pharmaceutical vendors. Therefore, the standard is supposed to be common enough to be used in many different use cases. Other aspects that lead to the adoption of FHIR as developer friendly standard are:

1. CDA can be found in many areas and pilot projects, but is often described as complicated, laborious and unwieldy in many use cases;

2. The complexity that HL7v2 provides due to its message structure;
3. The too high complexity of HL7v3.

FHIR provides a lot of improvements, as explained on the FHIR website:

- A strong focus on implementation: fast and easy to implement.
- Multiple implementation libraries, many examples available to kick-start development.
- Specification is free for use with no restrictions.
- Interoperability out-of-the-box: base resources can be used as is, but can also be adapted as needed for local requirements using Profiles, Extensions, Terminologies and more.
- Evolutionary development path from HL7 Version 2 and CDA: standards can co-exist and leverage each other.
- Strong foundation in Web standards: XML, JSON, HTTP, OAuth, etc.
- Support for RESTful architectures, seamless exchange of information using messages or documents, and service-based architectures.
- Concise and easily understood specifications.
- A human-readable serialization format for ease of use by developers.

The concept of "Profiles" in FHIR has been introduced to cope with great variety of needs in the different branches of healthcare, as it is not always mindful to use a resource in the exact the same way. A profile extends a resource by adding new properties or constrain a resource. Constraints may be specifications or modifications for datasets or multiplicity. Profiling is a tool to manage national characteristics, regulations and legislations. Furthermore, profiles can also describe the greater structure of FHIR applications, that is referencing.

FHIR resources use terminologies, which are comprehensive collections of medical terms like SNOMED CT or LOINC. FHIR REST-API defines the CRUD operations with similar HTTP requests, also using these URIs. Not only single resources can be acquired but bundles of connected resources like a certain CarePlan resource, defined by its ID, with all its activities. This works with the parameter "_include". FHIR resources can be accessed by GraphQL a query language for APIs and a runtime for fulfilling those queries with existing data and may be used as a facade service in front of a conformant RESTful API. There are many public testing FHIR servers available, that implement the entire FHIR specification, e.g. the HAPI FHIR server.

2.2 Synthetic Data Generation

It is difficult to find patient-level data of sufficient size for home care research projects, to be used in software development projects. Synthetic data has potential in those areas but much of the generated data is non-medical. For example, R and Python programming languages can generate non-medical datasets for supervised and unsupervised learning. Generative adversarial networks can also generate synthetic non-medical data.

In 2017 the Mitre Corporation developed the SyntheticMass project that emulated the healthcare data of the residents of Massachusetts. The data are fictitious, but realistic because they are based on the medical and demographic features of the residents of Massachusetts [13]. There are meanwhile a lot of instances of Synthea™, that generate huge amounts of synthetic medical records utilised in different medical use cases.

Synthea™ can be downloaded and operated with any computer terminal with simple commands and the use of Gradle. This permits a user to use their own modules to generate customized patient data (see Figure 1).

Synthea™ and the possible application to healthcare were mentioned in few articles of the medical literature. An article was published in 2019 that compared four quality measure standards e.g., colorectal cancer screening, with Synthea™ data and concluded that the demographics and services were accurately reflected in the data, but the health outcome measures did not accurately reflect state or national statistics [9].

Rossmiller used this application to generate realistic cancer patients for a nurse-facing app. In [6] Kaul used Synthea™ to model healthcare data in a TigerGraph graph database. The most recent (November 2020) article about Synthea™ was written by the originators of the SyntheticMass project and in the article they discussed synthetic COVID-19 data [14].

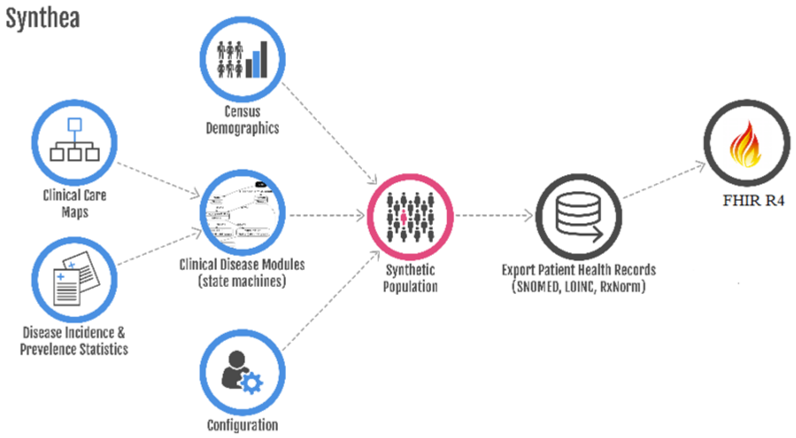

Synthea
Census Demographics
Clinical Care Maps
Clinical Disease Modules (state machines)
Disease Incidence & Prevelence Statistics
Configuration
Synthetic Population
Export Patient Health Records (SNOMED, LOINC, RxNorm)
FHIR R4

Figure 1. Overview of the Synthea framework

The datasets were based on three published studies and included 88 000+ COVID patients and 18 000+ hospitalizations. There was a 4.1 % mortality rate, 20.6 % hospitalization rate and 209 ventilator cases [14]. The longitudinal nature of the data is a huge plus for those studying the natural progression of diseases. Recurrent neural networks would be a favoured deep learning tool to mine that type of data. There are multiple years of data so some patients may have more than 100 test results recorded. On the other hand, some might want a cross-sectional view of the data, so the data would have to be re-configured. Synthea™ was selected in the INGE project to generate data in the home care sector generating QuestionnaireResponse resources. They are challenging resources, as their particular attributes are not predefined by the FHIR specification but governed by a Questionnaire FHIR resource defined in the context required in a project to gather feedback from users e.g. clinicians or patients.

3 USE CASE – HOME CARE CONSULTANCY

As mentioned above in the health and care sector machine learning has high potential to understand better risk factors of people in need of care and analyze the outcome of measures to improve, e.g., in-home care consultancy. However, data to learn from are needed. In the following we will describe the approach undertaken in the research project INGE integrate4care (digitale INtegrierte GEsundheits- und Pflegeversorgung mit IT-gestütztem Pflegeberatungsbesuch nach § 37.3 SGB XI/Digital integrated health and home care with IT-supported in-home care consultancy in conformance to § 37.3 social security statute book XI).

3.1 Background

In the care sector, particularly when it comes to home care provided by informal caregivers, data on care settings are hardly available. This is mainly due to lacking documentation of home care settings and even if, collected data are not standardized and often only available in print form. However, the majority of persons in need of care live at home and care is provided by informal caregivers. In Germany these are 80 % whereof 51 % are cared for by informal caregivers alone, i.e. without any support from care service providers. Informal caregivers are often lacking knowledge on how to provide care appropriately and efficiently though and feel left alone or even overburdened by their tasks. Beyond this, no care planning takes place for this group of people. To improve this situation quality assured in-home care consultancy is necessary.

In Germany in-home care consultancy is actually part of the health system and even required for informal caregivers receiving care allowances. The goal of the home visit is to verify that persons requiring care are appropriately taken care of at home and to advice the care recipient and their informal caregiver in regard to how the home care setting can be possibly improved or at least preserved. To achieve this, the health status of the person in need of care is crucial as well as informal caregivers perceiving their care tasks as manageable.

So far, no concrete standards are available for such in-home care consultancies and documentation of the assessment results is most of the times only available in print, if at all. In order to improve the quality of these consultancies and build a database for machine learning, the first prototype of an app was developed in course of the aforementioned nationally funded research project INGE that serves as an assessment tool for nurses to capture and document home care settings.

In order to receive an appropriate database for developing and testing machine learning algorithms, several mobile nursing services were and more will be recruited to use the INGE app during in-home consultancies. However, this will take time particularly since as mentioned above the COVID-19 pandemic caused a decrease in-home care consultancy. To start with developing machine learning algorithms already in the project before a suitable database could be built, synthetic generation of home care histories (health records) were used and literature reviewed for domain knowledge about possible risk predictors and suitable measures (interventions) to improve quality of data. As it turned out not much validated information is available about home care. Most of the research deals with key factors for institutionalization and possible preventive measures [3]. When it comes to risk assessment, usually clinical examinations are included [2] or knowledge about clinical data is required such as fluctuating blood glucose values that is not available during in-home care consulting. The reason is most likely that such risk assessment tools have been developed rather for professional care, e.g., the well-known NANDA-I nursing diagnosis. WHO provides measures to preserve a home care setting in their guidance on ‘Integrated care for older people. Guidelines on community-level interventions to manage declines in intrinsic capacity’. The level of evidence for these measures is rated in most cases as low or moderate though, so have to be treated with caution, but were considered for the INGE assessment and data model. Beyond this, domain knowledge was included in the INGE data model for synthetic data generation by distinguishing between care recipients and informal caregiver trajectories depending on whether a measure was applied or not in case the cut-off value for an assessment item was reached (see Section 3.3 for further details).

To further improve development of evidence-based domain knowledge as such as well as for synthetic data generation a visualization tool will be developed in INGE that allows to present health professionals such as geriatricians and nurses history graphs or other types of illustrations with real data collected during INGE in-home consultancies. This will allow health professionals to understand, e.g., correlations among key dimensions of home care and the impact of specific measures, so risk predictors and suitable measures can be identified. Implementing such findings in the INGE data model will allow to improve the risk assessment and quality of preventive measures suggested to nurses during an in-home care consultancy.

3.2 Development of the INGE App for in-Home Care Consultancy

The INGE-assessment was developed based on user requirements collected from target end-users by semi-structured interviews and established instruments to assess a home care setting (see Section 3.3 for more detail). During the development group-based expert walkthroughs were conducted with nurses, developers and usability experts to ensure usability and utility of the INGE app [4].

In the interviews the needs of informal caregivers, nurses experienced with in-home consulting and physicians from various medical subjects were collected to learn about the context of use of each target end-user group as well as required content and workflows. One of the main results was that in-home care consultancy takes place in form of a conversation rather than working through a predesigned assessment and it has a time limitation. That means the INGE app needs to allow for flexible and direct access to assessment items to document a conversation and the number of assessment items needs to be limited and much focused on what is relevant for home care.

Besides this, all health professionals were in favor of systematic capturing and documentation of a home care setting to allow for suitable care planning. An important part of care planning involves suggesting measures in case an issue has been identified in the home care setting. This may involve bathroom tools, a walker or day care to disburden the informal caregiver. Following-up whether the measure was taken on or there is a problem, e.g. the statutory health insurance has not approved on it or it did not turn out to be practicable was considered important by nurses to ensure effective care planning.

3.3 The INGE Assessment Instrument

To ensure systematic capturing of a home care setting two validated and in Germany well established assessment instruments were used as base for the development of a home care assessment in INGE, namely the NBA (Neues Begutachtungsassessment zur Feststellung der Pflegebedürftigkeit/New Assessment Tool for determining dependency on care) [15] and the BICS-D (Berliner Inventar zur Angehörigenbelastung-Demenz/Berlin Inventory of Caregiver Stress-Dementia) (BICS-D). The BICS-D [16, 11] which is used to assess psychosocial impairments and burden that arise as a result of caregiving.

The NBA was developed for home visits by the Medical Service of the German statutory health insurance (Medizinischer Dienst der Krankenkassen (MDK)). Main goal of these visits is to assess the level of care needed by a care recipient making the recipient and his or her informal caregiver eligible for care services funded by their statutory health insurance. The NBA consists of eight dimensions: mobility, behavior and psychological problems, cognitive and communication abilities, selfcare, ability to deal with illness-/therapy-related demands and stress, managing everyday life and social contacts as well as activities outside of the house and household maintenance. For each dimension the limitation of independence or abilities is summarized on a five-point scale (e.g. 1 = independent/no limitations, 2 = minor limitations [of independence or abilities], 3 = significant limitations [of independence or abilities], 4 = severe limitations [of independence or abilities], 5 = complete/extensive loss [of independence or abilities]). The sum values of the first six modules are weighted, summed and then used to determine the level of care needed [15].

Of the NBA five dimensions (35 items) were used and slightly changed in wording for the INGE assessment mainly to shorten and facilitate scanning of dimension topics in the app by nurses:

1. Cognition and Communication (7 items, four-point scale, 1 = abilities present, 2 = abilities predominantly present, 3 = abilities present to small extent, 4 = abilities not present, e.g. 'How is the person's orientation in relation to time and place?')
2. Mobility (5 items, four-point scale, 1 = independent, 2 = predominantly independent, 3 = predominantly non-independent, 4 = non-independent, e.g. 'How independently can the person move or change the position of his or her body?')
3. Self-Care (12 items, four-point scale, 1 = independent, 2 = predominantly independent, 3 = predominantly non-independent, 4 = non-independent, e.g. 'How independently can the person take care of himself/herself in regard to eating and drinking?')
4. Behavior (8 items, five-point scale, 1 = never, 2 = rarely [less than once a week], 3 = sometimes [about once a week], 4 = often [several times a week but not every day], 5 = very often [once or several times a day], e.g. 'How often does the person need assistance because of anxieties?')
5. Social environment (3 items, four-point scale, 1 = independent, 2 = predominantly independent, 3 = predominantly non-independent, 4 = non-independent, e.g. 'How independently can the person still arrange and plan their daily schedule?')

The response option 'unclear' was added to each scale. Based on the values of the NBA and adjusted to the new instrument, a sum value is calculated for each individual dimension and cut-off values based on this indicate a certain status visualized in a traffic light system with:

green: independent or no limitations,

yellow: minor limitations [of independence or abilities],

orange: significant limitations [of independence or abilities],

red: severe limitations [of independence or abilities] to complete/extensive loss [of independence or abilities].

The BICS-D was developed as multidimensional measure of informal caregiver burden in home-care settings. The instrument is sensitive to changes and makes it possible to target interventions (Schlomann et al., 2020). A short version of the BICS-D [10] was mainly used as base for INGE measures in regard to informal caregiver burden. It includes six dimensions: burden due to cognitive losses, burden due to behavioral changes, personal constraints, practical caregiving tasks, lack of social recognition and accepting coping. Based on the informal caregivers' BICS-D-score typical risks of the caregiving situation can be assessed, namely the risk of depression for informal caregivers, the risk of violence in care and the risk of institutionalization of the care recipient [10]. Dimension ‘negative appraisal of one’s own care-giving’ was additionally chosen from the long version as part of INGE because it has proven to be one of the biggest influencing factors on the probability of institutionalization [8].

Of the BIZA-D 3 dimensions (11 items) were chosen and slightly changed in wording for the INGE assessment:

1. Personal constraints (5 items, five-point Likert scale, 1 = never, 5 = always, e.g. ‘Do you feel that you have too little time for hobbies/interests?’)
2. Burden due to cognitive losses (3 Items, five-point Likert scale, 1 = not at all, 5 = very much, e.g. ‘The person affected is no longer accessible to logical arguments. – if ‘yes’ – ‘How much of a burden is this on you?’)
3. Low self-evaluation (3 items, five-point Likert scale, 1 = never, 5 = always, e.g. ‘Do you have the feeling that you are making mistakes when providing care?’)

Based on the results the informal caregiver’s risk for 3 different outcomes (developing depression, violence between care recipient and caregiver, institutionalization of the care recipient) is calculated for each dimension and presented in a traffic light system:

green: low risk,

yellow: moderate risk,

orange: high risk,

red: extreme risk.

Beyond the assessment as described above the INGE app includes a catalogue with about 200 measures, as suggested by the Remedies and Aid Registry for care in Germany. These are aids to, e.g., support activities of daily living, care services such as day care or clinical examinations like gerontological-psychiatric examination. To facilitate selection of a suitable measurement a cut-off value was defined for each assessment item and in case this value was reached, a potential risk is assumed and three possibly suitable measures are presented in the app. The cut-off value as well as the preselected measures were defined by the nurses in the INGE-project on base of their extensive experience with in-home care consultancy, however, nurses are always free to select another measure from the catalogue or decide not to suggest a measure at all. The INGE machine learning model will analyze the decisions of nurses in correlation with other assessment ratings and active measures to improve the preselection and evaluate suitability of the cut-off values. Information on personal data of the care recipient, in particular the health status (medication, diagnoses, hospital stays etc.) and personal data of the informal caregiver will be considered for analysis as well.

3.4 Translating the INGE Data Model into FHIR

The objects identified in the user requirements phase were mapped to FHIR resources. Some custom extensions have been defined to store additional data that are not in the default FHIR specification but are necessary for the INGE project. Figure 2 describes the INGE data model with all utilised FHIR resources.

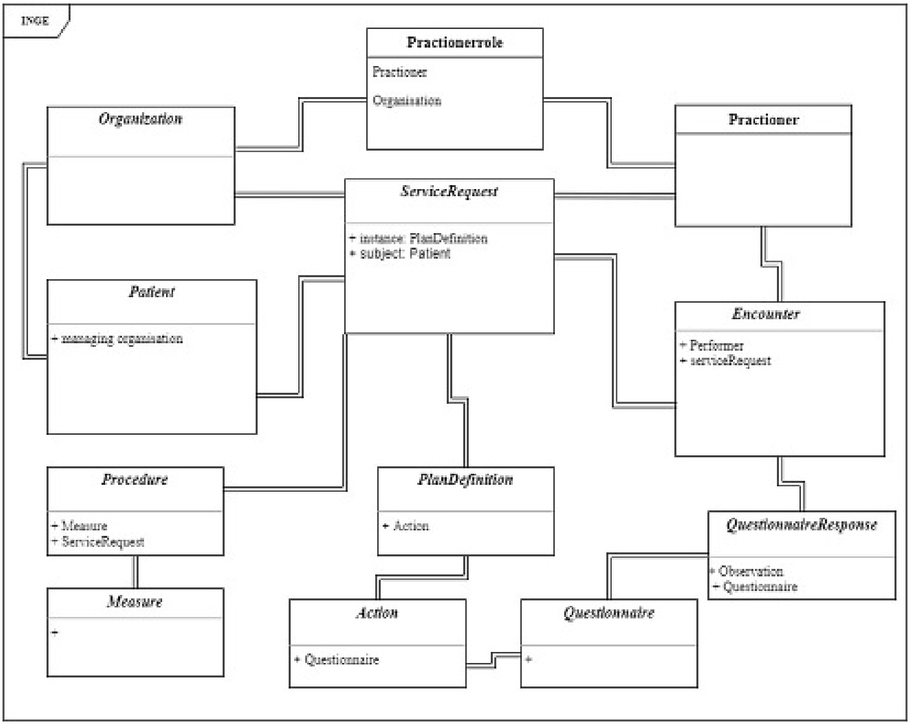
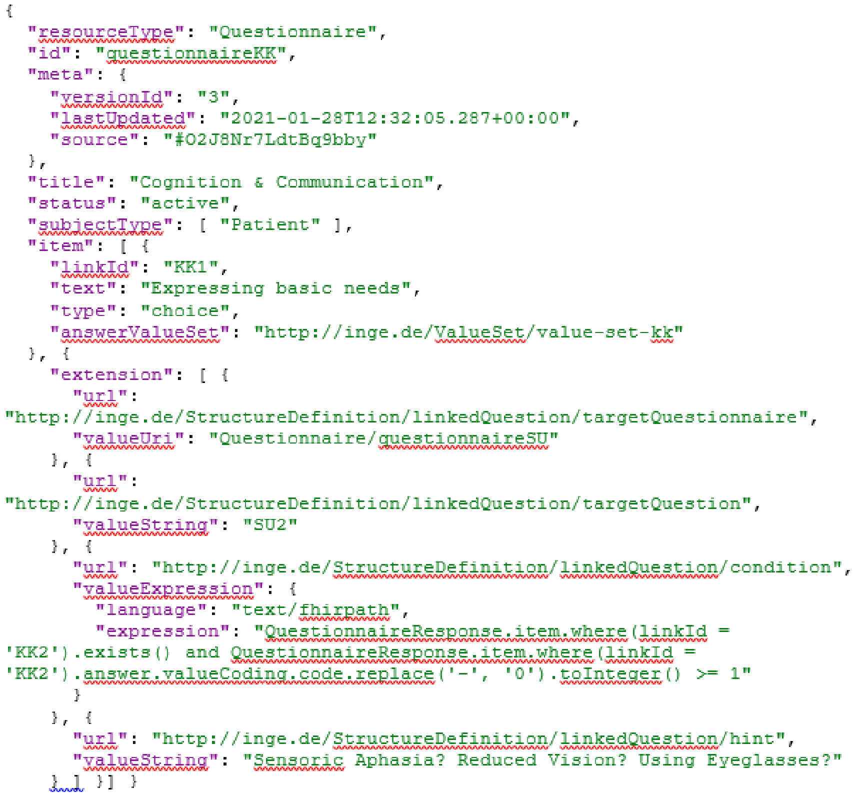


Figure 2. INGE data model

The core resources of the INGE assessment were mapped to "Questionnaire" and "QuestionnaireResponse" resources. Most of the question items in the assessments are of type 'choice' and shares the same list of options. Some ValueSet resources has been created for these repeating lists of options. Questionnaires and forms permeate healthcare. They are used to capture administrative data, claims data, clinical information, research information, for public health reporting – every type of data that is manipulated by healthcare systems. They provide a user-friendly mechanism for capturing data in a consistent way. Here is an example FHIR Questionnaire resource in JSON format.

In FHIR, forms are represented using the Questionnaire resource and completed forms are represented using the QuestionnaireResponse resource. The base FHIR specification defines these resources but does not provide much guidance around how they should be used, nor does it set minimal expectations for interoperation. This implementation guide provides a set of guidance around the use of Questionnaire and QuestionnaireResponse. Specifically, it provides answers to – and conformance expectations around – questions such as those in Figure 3.

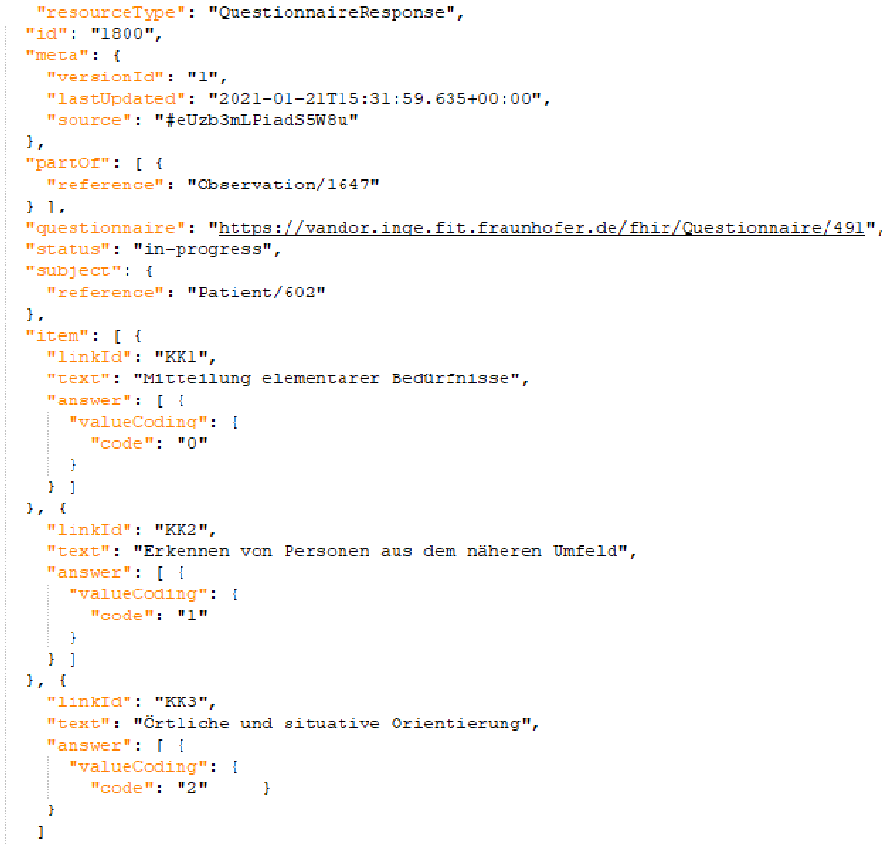


```
{
  "resourceType": "Questionnaire",
  "id": "questionnaireKK",
  "meta": {
    "versionId": "3",
    "lastUpdated": "2021-01-28T12:32:05.287+00:00",
    "source": "#O2J8Nr7LdtBq9bby"
  },
  "title": "Cognition & Communication",
  "status": "active",
  "subjectType": [ "Patient" ],
  "item": [ {
    "linkId": "KK1",
    "text": "Expressing basic needs",
    "type": "choice",
    "answerValueSet": "http://inge.de/ValueSet/value-set-kk"
  }, {
    "extension": [ {
      "url":
"http://inge.de/StructureDefinition/linkedQuestion/targetQuestionnaire",
      "valueUri": "Questionnaire/questionnaireSU"
    }, {
      "url":
"http://inge.de/StructureDefinition/linkedQuestion/targetQuestion",
      "valueString": "SU2"
    }, {
      "url": "http://inge.de/StructureDefinition/linkedQuestion/condition",
      "valueExpression": {
        "language": "text/fhirpath",
        "expression": "QuestionnaireResponse.item.where(linkId =
'KK2').exists() and QuestionnaireResponse.item.where(linkId =
'KK2').answer.valueCoding.code.replace('-', '0').toInteger() >= 1"
      }
    }, {
      "url": "http://inge.de/StructureDefinition/linkedQuestion/hint",
      "valueString": "Sensoric Aphasia? Reduced Vision? Using Eyeglasses?"
    } ] }] }
```

Figure 3. Questionnaire instance

The resource QuestionnaireResponse in FHIR is an instance based always on an existing Questionnaire resource, so the relation between both resources is one Questionnaire to many QuestionnaireResponses. This can be seen similar to class and object in object oriented programming, so the Questionnaire resource is the class and the QuestionnaireResponse is the object. In Figure 4 you can see an instance of a QuestionnaireResponse based on the above in Figure 3 shown Questionnaire instance.

“Patient” was as well a key FHIR resource to store data like ID, birthdate, gender etc.. ServiceRequest Observations and procedures are based on service re-



```
 "resourceType": "QuestionnaireResponse",
"id": "1800",
"meta": {
  "versionId": "1",
  "lastUpdated": "2021-01-21T15:31:59.635+00:00",
  "source": "#eUzb3mLPiadS5W8u"
},
"partOf": [ {
  "reference": "Observation/1647"
} ],
"questionnaire": "https://vandor.inge.fit.fraunhofer.de/fhir/Questionnaire/491",
"status": "in-progress",
"subject": {
  "reference": "Patient/602"
},
"item": [ {
  "linkId": "KK1",
  "text": "Mitteilung elementarer Bedürfnisse",
  "answer": [ {
    "valueCoding": {
      "code": "0"
    }
  } ]
}, {
  "linkId": "KK2",
  "text": "Erkennen von Personen aus dem näheren Umfeld",
  "answer": [ {
    "valueCoding": {
      "code": "1"
    }
  } ]
}, {
  "linkId": "KK3",
  "text": "Örtliche und situative Orientierung",
  "answer": [ {
    "valueCoding": {
      "code": "2"     }
  }
 ]
```

Figure 4. QuestionnaireResponse instance

quests. This resource type is mainly used for storing the PlanDefinition related to an Observation, also the status of a Procedure.

3.5 Synthetical Data Generation

As described above, we had at the start date of the INGE project no related data in digital form. The data documented on paper was so unstructured that made it almost impossible to convert it in digital form according to the data model of INGE. A decision was made in the INGE project to use the SyntheaTM synthetic patient population Simulator to output synthetic, realistic (but not real) patient data and associated health records for the identified home care setting. The INGE project adopted the international version of SyntheaTM and added the "Questionnaire" and "QuestionnaireResponse" resources to the code base of SyntheaTM. In addition to the default implementation, this enhancement version includes the following:

- Modified Disease Module (Condition), to generate ICD-10 conditions.
- Generate data in alternative geographic locations with relevant geographical standard, such as Europe.
- Generate Questionnaires, QuestionnairesResponses type of FHIR R4 resources in a flexible and configurable way.

3.6 Machine Learning Model

There are many projects utilising generated data by SyntheaTM in machine learning projects e.g. “FHIR from Jupyter”, where FHIR resources are impoted into Jupyter notebooks through the FHIR API, FHIR in Spark and then used in predictive modeling. The Bunsen project of Cerner lets users load, transform and analyze FHIR data with Apache Spark. It offers Java and Python APIs to convert FHIR resources into Apache Spark Datasets, which then can be explored with the full power of that platform, including with Spark SQL. In the INGE project we have followed the Bunsen framework approach and loaded generated data by SyntheaTM into Jupyter notebooks and processed the data with Apache Spark utilizing different machine learning algorithms (see Figures 7 and 8).

As shown in Figure 3, we combine generated data with real world data by using anonymised real-world data as the zero patients in the synthetical data generator. The generator implements as well further rules to ensure that the generated data are realistic to the most possible extent.

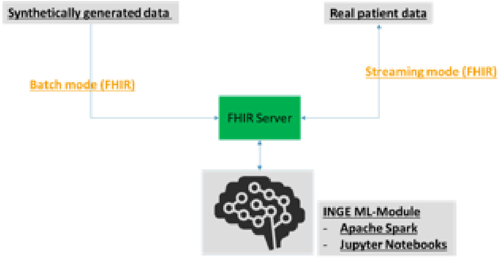


Figure 5. INGE machine learning approach

The generated data are as explained above imported into Jupyter notebooks, where they are loaded and processed. The experts can visualize the data through different diagram types as shown in Figure 4, where data of all patients for a specific questionnaire are displayed or in Figure 5, where recommended measures (interventions) by the advisory nurses over many years are being shown.

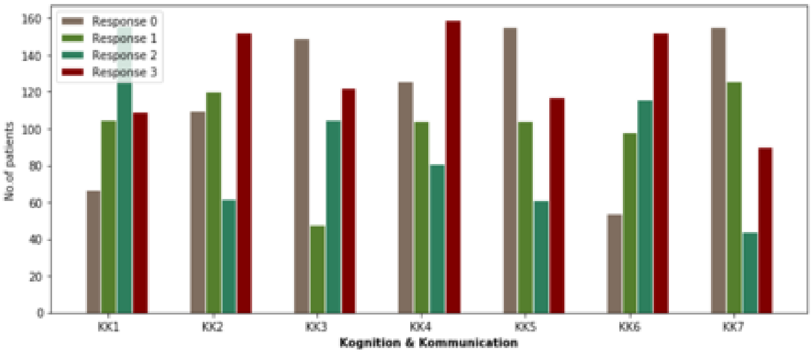


Figure 6. Visualization of questionnaire responses

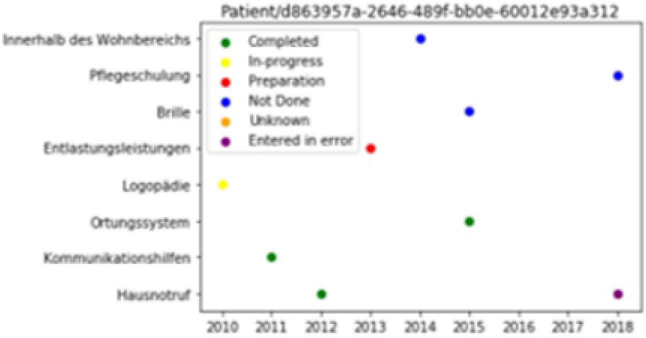


Figure 7. Visualization of measures ongoing for a patient over years

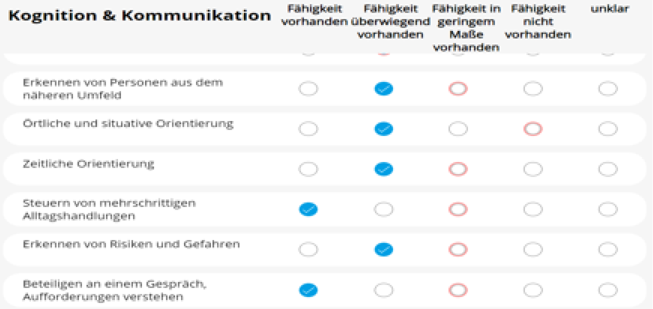
Based on the loaded FHIR resources two sets of ML models were implemented:

Model 1: To identify and suggest measures based on each questionnaire response.

Model 2: To learn patterns in the questionnaire responses filled out by the patients over time to make more accurate predictions in future.

Based on the FHIR resources in the use case of the INGE project, we have built different Machine Learning Models as follows:

- Multi-Class Classification problem needs to be formulated,
- FHIR Questionnaire Response Resource is used as input,
- Data is transformed using One hot encoder estimator and converted to a vector input,
- 70/30 ratio of data set is used for training and testing,

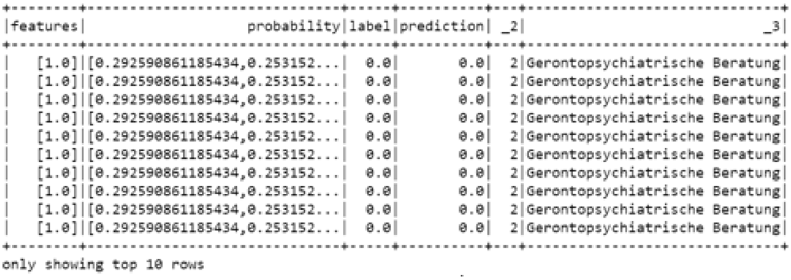


Kognition & Kommunikation	Fähigkeit vorhanden	Fähigkeit überwiegend vorhanden	Fähigkeit in geringem Maße vorhanden	Fähigkeit nicht vorhanden	unklar
Erkennen von Personen aus dem näheren Umfeld	○	☑	○	○	○
Örtliche und situative Orientierung	○	☑	○	○	○
Zeitliche Orientierung	○	☑	○	○	○
Steuern von mehrschrittigen Alltagshandlungen	☑	○	○	○	○
Erkennen von Risiken und Gefahren	○	☑	○	○	○
Beteiligen an einem Gespräch, Aufforderungen verstehen	☑	○	○	○	○

Figure 8. Questionnaire set for which the two ML models are implemented

- Various Algorithms used:
  - Logistic Regression,
  - Random Forest,
  - Decision Tree Classifier,
  - Naive Bayes Classifier,
- Prediction is evaluated based on accuracy metrics.

  Test Results for ML Model 1 (see Figure 9).



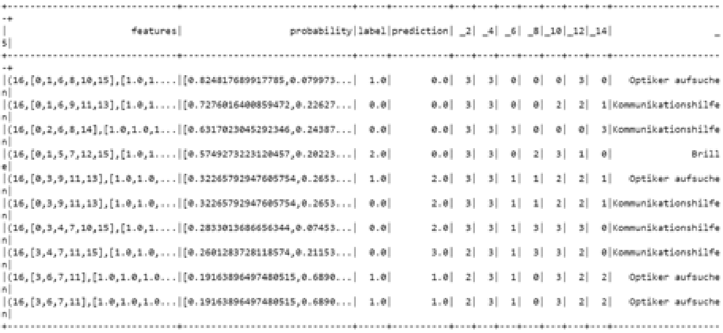
```
+--------+------------------------------+-----+----------+---+------------------------------+
|features|                   probability|label|prediction| _2|                            _3|
+--------+------------------------------+-----+----------+---+------------------------------+
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
|   [1.0]|[0.292590861185434,0.253152...|  0.0|       0.0|  2|Gerontopsychiatrische Beratung|
+--------+------------------------------+-----+----------+---+------------------------------+
only showing top 10 rows
```

Figure 9. ML model results predicting measures for each questionnaire item

Test Results for ML Model 2 (see Figure 10).

4 CONCLUSIONS AND FUTURE RESEARCH AREAS

In this paper we have presented a disruptive approach of building a medical data model in the home care sector, generating synthetical data to overcome lack of data, privacy, and legal issues. We described the current state of medical data formats (FHIR) and the framework Synthea™ for generation of simulated health record data. To increase the benefits, this work requires a lot of effort to ensure



```
+----------------------------+-----------------------------+-----+----------+---+---+---+---+---+---+---+-------------------
-+
|                    features|                  probability|label|prediction| _2| _4| _6| _8|_10|_12|_14|                  _
5|
+----------------------------+-----------------------------+-----+----------+---+---+---+---+---+---+---+-------------------
-+
|(16,[0,1,6,8,10,15],[1.0,1....|[0.8248176899177785,0.079973...|  1.0|       0.0|  3|  3|  0|  0|  0|  3|  0|   Optiker aufsuche
n|
|(16,[0,1,6,9,11,13],[1.0,1....|[0.7276016400859472,0.22627...|  0.0|       0.0|  3|  3|  0|  0|  2|  2|  1|Kommunikationshilfe
n|
|(16,[0,2,6,8,14],[1.0,1.0,1...|[0.6317023045292346,0.24387...|  0.0|       0.0|  3|  3|  3|  0|  0|  0|  3|Kommunikationshilfe
n|
|(16,[0,1,5,7,12,15],[1.0,1....|[0.5749273223120457,0.20223...|  2.0|       0.0|  3|  3|  0|  2|  3|  1|  0|              Brill
e|
|(16,[0,3,9,11,13],[1.0,1.0,...|[0.32265792947605754,0.2653...|  1.0|       2.0|  3|  3|  1|  1|  2|  2|  1|   Optiker aufsuche
n|
|(16,[0,3,9,11,13],[1.0,1.0,...|[0.32265792947605754,0.2653...|  0.0|       2.0|  3|  3|  1|  1|  2|  2|  1|Kommunikationshilfe
n|
|(16,[0,3,4,7,10,15],[1.0,1....|[0.2833013686656344,0.07453...|  0.0|       2.0|  3|  3|  1|  3|  3|  3|  0|Kommunikationshilfe
n|
|(16,[3,4,7,11,15],[1.0,1.0,...|[0.2601283728118574,0.21153...|  0.0|       3.0|  2|  3|  1|  3|  3|  2|  0|Kommunikationshilfe
n|
|(16,[3,6,7,11],[1.0,1.0,1.0...|[0.19163896497480515,0.6890...|  1.0|       1.0|  2|  3|  1|  0|  3|  2|  2|   Optiker aufsuche
n|
|(16,[3,6,7,11],[1.0,1.0,1.0...|[0.19163896497480515,0.6890...|  1.0|       1.0|  2|  3|  1|  0|  3|  2|  2|   Optiker aufsuche
n|
+----------------------------+-----------------------------+-----+----------+---+---+---+---+---+---+---+-------------------
```

Figure 10. ML model results to learn patterns in the questionnaire responses and then predicting measures (interventions)

the quality of data. One obstacle is that not much evidence-based knowledge for this domain is available in literature and suitable data repositories do not exist yet. Therefore, domain knowledge from literature as available and health professionals in the project was included in the INGE data model and will be improved constantly by analysing real data collected from in-home consultancies conducted in course of INGE. Beyond this, data from these consultancies will be visualized to health professionals for further analysis and enhancement of domain knowledge.

Acknowledgments

This work was partially funded by the European Union from the European Regional Development Fund (EFRE) and the State of Rhineland-Westphalia (NRW), Germany. The authors would like to acknowledge the support of the INGE consortium.

REFERENCES

[1] Chen, J.—Chun, D.—Patel, M.—Chiang, E.—James, J.: The Validity of Synthetic Clinical Data: A Validation Study of a Leading Synthetic Data Generator (Synthea) Using Clinical Quality Measures. BMC Medical Informatics and Decision Making, Vol. 19, 2019, Art. No. 44, doi: 10.1186/s12911-019-0793-0.

[2] Culo, S.: Risk Assessment and Intervention for Vulnerable Older Adults. British Columbia Medical Journal, Vol. 53, 2011, No. 8, pp. 421–425.

[3] de Almeida Mello, J.—Declercq, A.—Cès, S.—Van Durme, T.—Van Audenhove, C.—Macq, J.: Exploring Home Care Interventions for Frail Older People in Belgium: A Comparative Effectiveness Study. Journal of the American Geriatrics Society, Vol. 64, 2016, No. 11, pp. 2251–2256, doi: 10.1111/jgs.14410.

[4] Gappa, H.—Nordbrock, G.—Mohamad, Y.—Velasco, C. A.: Group-Based Expert Walkthroughs to Compensate for Limited Access to Target User Groups as in Case of Chronically Ill Patients. In: Miesenberger, K., Kouroupetroglou, G., (Eds.): Computers Helping People with Special Needs (ICCHP 2018). Springer, Cham, Lecture Notes in Computer Science, Vol. 10896, 2018, pp. 71–74, doi: 10.1007/978-3-319-94277-3_13.

[5] Walonoski, J.—Kramer, M.—Nichols, J.—Quina, A.—Moesel, C.—Hall, D.—Duffett, C.—Dube, K.—Gallagher, T.—McLachlan, S.: Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record. Journal of the American Medical Informatics Association, Vol. 25, 2018, No. 3, pp. 230–238, doi: 10.1093/jamia/ocx079.

[6] Kaul, A.: Modeling Healthcare Data with Graph Databases Using TigerGraph and Synthea to Create a Synthetic Healthcare System. July 6, 2020, https://towardsdatascience.com/modeling-healthcare-data-with-graph-databases-3e3695bcae3c.

[7] Braunstein, M. L.: Patient-Physician Collaboration on FHIR (Fast Healthcare Interoperability Resources). 2015 International Conference on Collaboration Technologies and Systems (CTS), IEEE, 2015, pp. 501–503, doi: 10.1109/CTS.2015.7210457.

[8] Oliva y Hausmann, A.—Schacke, C.—Zank, S.: Caregiving Relatives of People with Dementia: Which Factors Influence the Transfer from Home-Care to Nursing Home? Psychotherapie, Psychosomatik und Medizinische Psychologie, Vol. 62, 2012, Nos. 9–10, pp. 367–374, doi: 10.1055/s-0032-1323708 (in German).

[9] Rossmiller, R.: Synthetic Health Data Generation: My First Experience with Synthea. July 31, 2019, https://medium.com/@rrossmiller24/my-first-experience-with-synthea-synthetic-health-data-generation-74fffd74a138.

[10] Schacke, C.—Zank, S.: Das Berliner Inventar zur Angehörigenbelastung – Demenz (BIZA-D). Manual für die Praxisversion (BIZA-D-PV). ZPE-Schriftenreihe, 23, Siegen, Zentrum für Planung und Evaluation Sozialer Dienste der Universität Siegen, 2009 (in German).

[11] Schlomann, A.—Schacke, C.—Leipold, B.—Zank, S.: Berlin Inventory of Caregiver Stress – Dementia (BICS-D). The Gerontologist, Vol. 61, 2021, No. 5, pp. e173–e184, doi: 10.1093/geront/gnz195.

[12] Benson, T.—Grieve, G.: Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. 3rd Edition. Springer, London, Health Information Technology Standards, 2016, doi: 10.1007/978-3-319-30370-3.

[13] Walonoski, J.—Kramer, M.—Nichols, J. et al.: Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record. Journal of the American Medical Informatics Association, Vol. 25, 2018, No. 3, pp. 230–238, doi: 10.1093/jamia/ocx079.

[14] Walonoski, J.—Klaus, S.—Granger, E. et al.: Synthea™ Novel Coronavirus (COVID-19) Model and Synthetic Data Set. Intelligence-Based Medicine, Vols. 1–2, 2020, Art. No. 100007, doi: 10.1016/j.ibmed.2020.100007.

[15] Wingenfeld, K.—Büscher, A.—Gansweid, B.: Das Neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Abschlussbericht zur Hauptphase 1,

Überarbeitete/Korrigierte Fassung, Bielefeld/Münster, 2008 (in German).

[16] Zank, S.—Schacke, C.—Leipold, B.: Längsschnittstudie zur Belastung Pflegender Angehöriger von demenziell Erkrankten (LEANDER): Ergebnisse der Evaluation von Entlastungsangeboten. Zeitschrift für Gerontopsychologie und -psychiatrie, Vol. 20, 2007, No. 4, pp. 239–255, doi: 10.1024/1011-6877.20.4.239.

[17] HL7 FHIR. `http://hl7.org/fhir/`.

Yehya Mohamad holds his Ph.D. in computer science from the Technical University of Aachen (Germany). His Ph.D. works focused on vital sign detection in real time from sensors and wearables. He holds a Masters in computer science from the Technische Universität Berlin. He is a senior researcher, lecturer, and digital health expert at the Fraunhofer Institute for Applied Information Technology FIT (Germany) since 2001. His research interests include digital health, compliance, usability and accessibility of information systems, standards, user profiling, internet technologies, and web of things. He was involved in several funded European projects and published about hundred research papers in prestigious conferences and journals. He also organized several successful international research events.

Alexander Gabber received his Bachelor of Arts in sociology and psychology from the Friedrich-Schiller-University in Jena, Germany in 2016 and his Master of Science in gerontology from the Friedrich-Alexander-University in Nuremberg, Germany in 2019. Currently he is a research associate and Ph.D. student at the Department of Special Education and Rehabilitation at the University of Cologne, Germany. His research interests are usability of mHealth and eHealth applications for healthcare staff and the elderly.

Sonja Heidenblut is Psychologist and received her Ph.D. from the University of Cologne in 2012. Currently she is a research associate at the Chair of Gerontology at the University of Cologne. Her research interests are gerontology and geriatric assessment.

Daniel Zenz studied business economics in Bochum and Multimedia Producer in Cologne. He founded smart-Q Softwaresysteme GmbH in 2010 with now 25 employees. The business and scientific focus is in the research and creation of user friendly highly specialized medical software such as patient health records. The invented products focus all on the medical market and are highly connected via interfaces with other medical applications. The highly secured web based products are used by over 5 000 doctors and nurses in Germany.

Anam Siddiqi is M.Sc. computer science student at Rheinische Friedrich-Wilhelms-Universität Bonn. Working as Student Assistant at the Fraunhofer Institute for Applied Information Technology FIT since 2 years 9 months.

Henrike Gappa has her Masters degree in special education (Lesley College, Cambridge, USA, 1984) and the first state examination for teachers of children with special needs (University of Cologne, 1986). Since 1986 she has worked in different research organizations such as the University of Cologne with focus on computer-aided intervention for people with special needs. At the Fraunhofer Institute for Applied Information Technology FIT, she is currently a senior researcher involved in several national and European research projects related to Web accessibility, usability and e-health. Her key working areas are user requirements engineering and design of multimodal user interfaces for people with disabilities and elder persons.