STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS 287

Applying the FAIR Principles to Accelerate Health Research in Europe in the post COVID-19 Era

Proceedings of the 2021 EFMI Special Topic Conference

22nd - 24th NOVEMBER, 2021

APPLYING THE FAIR PRINCIPLES TO ACCELERATE HEALTH RESEARCH IN EUROPE IN THE POST-COVID19 ERA

Health Research Data Ecosystem

Special Topic Conference

STC 2021

Informatics for Discoverability and Trust in Biomedical Research

EFMI co-organised by SOCIEDAD ESPAÑOLA DE SEIS INFORMÁTICA DE LA SALUD IBiS INSTITUTO DE BIOMEDICINA DE SEVILLA

Editors: Jaime Delgado
Arriel Benis
Paula de Toledo
Parisis Gallos
Mauro Giacomini
Alicia Martínez-García
Dario Salvi

IOS Press

Medical Informatics has increasingly come into focus in the last couple of years, as the importance of managing and interpreting health data in dealing with a global pandemic has become dramatically apparent.

This book presents the proceedings of the 2021 European Federation for Medical Informatics (EFMI) Special Topic Conference (STC), originally planned as a live event in Seville, Spain, but ultimately held as a virtual event from 22 – 24 November 2021. This conference focused on applying the FAIR principles (Findability, Accessibility, Interoperability and Reusability) to accelerate health research in Europe in the post COVID-19 era. The 38 papers included here are divided into 5 sections, and topics covered include: methods for the adoption of FAIR principles; FAIR-based precision medicine; AI in FAIR data-driven health; privacy and security aspects of applying FAIR in health research; FAIR and infectious-disease research data (including Covid-19); FAIR in infrastructures and software; metadata, ontologies and terminologies to support the sharing of health research data; and paradigms for sharing health research data.

Offering a state-of-the-art overview of medical informatics in the post-Covid era, the book will be of interest to all those working in the field.

ISBN: 978-1-64368-236-5
9 781643 682365

ISBN 978-1-64368-236-5 (print)
ISBN 978-1-64368-237-2 (online)
ISSN 0926-9630 (print)
ISSN 1879-8365 (online)

APPLYING THE FAIR PRINCIPLES TO ACCELERATE HEALTH RESEARCH IN EUROPE IN THE POST COVID-19 ERA

Studies in Health Technology and Informatics

International health informatics is driven by developments in biomedical technologies and medical informatics research that are advancing in parallel and form one integrated world of information and communication media and result in massive amounts of health data. These components include genomics and precision medicine, machine learning, translational informatics, intelligent systems for clinicians and patients, mobile health applications, data-driven telecommunication and rehabilitative technology, sensors, intelligent home technology, EHR and patient-controlled data, and Internet of Things.

Studies in Health Technology and Informatics (HTI) series was started in 1990 in collaboration with EU programmes that preceded the Horizon 2020 to promote biomedical and health informatics research. It has developed into a highly visible global platform for the dissemination of original research in this field, containing more than 250 volumes of high-quality works from all over the world.

The international Editorial Board selects publications with relevance and quality for the field. All contributions to the volumes in the series are peer reviewed.

Volumes in the HTI series are submitted for indexing by MEDLINE/PubMed; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Google Scholar; Scopus; EMCare.

Volume 287

Recently published in this series

Vol. 286 R. Marcilly, L. Dusseljee-Peute, C.E. Kuziemsky, X. Zhu, P. Elkin and C. Nøhr (Eds.), Context Sensitive Health Informatics: The Role of Informatics in Global Pandemics

Vol. 285 B. Blobel and M. Giacomini (Eds.), pHealth 2021 – Proceedings of the 18th International Conference on Wearable Micro and Nano Technologies for Personalized Health – 8–10 November 2021, Genoa, Italy

Vol. 284 M. Honey, C. Ronquillo, T.-T. Lee and L. Westbrooke (Eds.), Nurses and Midwives in the Digital Age – Selected Papers, Posters and Panels from the 15th International Congress in Nursing Informatics

ISSN 0926-9630 (print)
ISSN 1879-8365 (online)

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era

Proceedings of the 2021 EFMI Special Topic Conference

Edited by

Jaime Delgado

Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Arriel Benis

HIT – Holon Institute of Technology, Holon, Israel

Paula de Toledo

Universidad Carlos III, Madrid, Spain

Parisis Gallos

Health Informatics Laboratory, National and Kapodistrian University of Athens, Greece

Mauro Giacomini

University of Genoa, Genova, Italy

Alicia Martínez-García

IBiS – Institute of Biomedicine of Seville, Sevilla, Spain

and

Dario Salvi

Malmö University, Malmö, Sweden

IOS Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-64368-236-5 (print)
ISBN 978-1-64368-237-2 (online)
Library of Congress Control Number: 2021950401
doi: 10.3233/SHTI287

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

PRINTED IN THE NETHERLANDS

Preface

This volume presents the proceedings of the 2021 EFMI Special Topic Conference (STC) organized in November 2021 as a virtual conference. This conference focuses on applying the FAIR principles to accelerate health research in Europe in the post COVID-19 era. The conference invited paper submissions, in particular those related to the following topics:

- Methods for the adoption of FAIR principles
- FAIR-based precision medicine
- Artificial Intelligence in FAIR-data driven health
- Privacy and security aspects of applying FAIR in health research
- FAIR and Covid-19 (and other infectious diseases) research data
- FAIR for infrastructures and software
- Metadata, ontologies and terminologies to support the sharing of health research data
- Paradigms for sharing health research data.

All the papers in this book of proceedings received the highest marks in the peer review process, and the volume is organized into several sections. The most popular tracks among the authors were those on Metadata, Methods and Artificial Intelligence, and cover a wide area of applications. The remaining papers fall into the categories of Data and Experiences. As expected, many papers focus on FAIR and COVID-19.

STC 2021 was initially planned as a face-to-face event, to be held in Seville, Spain, and organized by the IBiS (Institute of Biomedicine of Seville) and the SEIS ("Sociedad Española de Informática de la Salud"), the Spanish representative in EFMI. However, due to the situation and travel restrictions with regard to the Covid pandemic, the conference was conducted online.

The Scientific Program Committee (SPC) included representatives from a number of EFMI Working Groups and the EFMI Board, as well as independent experts. The SPC consisted of the following: Jaime Delgado (Chair), Arriel Benis, Paula de Toledo, Parisis Gallos, Mauro Giacomini, Alicia Martínez-García and Dario Salvi.

On behalf of the Scientific Program Committee, I would first like to warmly thank all the authors who submitted their papers to the conference. Many thanks are also due to the reviewers, whose voluntary work contributed to the quality of the conference, not forgetting the scientific program committee itself for putting the whole conference together through its meetings and individual work.

Jaime Delgado
Chair of Scientific Programme Committee
October 2021

This page intentionally left blank

EFMI Special Topic Conference 2021 Scientific Programme Committee and Reviewers

Jaime Delgado, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
Alicia Martínez-García, IBiS - Institute of Biomedicine of Seville, Sevilla, Spain
Arriel Benis, HIT - Holon Institute of Technology, Holon, Israel
Dario Salvi, Malmö University, Malmö, Sweden
Mauro Giacomini, University of Genoa, Genova, Italy
Parisis Gallos, National and Kapodistrian University of Athens, Greece
Paula de Toledo, Universidad Carlos III, Madrid, Spain

Leila Ahmadian
Yasser Alsafadi
Alirıza Arıbaş
Mansoor Baig
Alireza Banaye Yazdipour
Mohamed Ben Said
Elena Bernad
Oana Sorina Chirila
Mihaela Crișan - Vida
Kerstin Denecke
Martin Dugas
Peter Elkin
Amado Espinosa
Mircea Focsa
Jan Gaebel
Denise Giles
Natalia Grabar
Kemal Hakan Gulkesen
Angelika Händel
Kristiina Häyrinen
Jacob Hofdijk
Lukas Huber
Ursula Hübner
Sanja Ivankovic
Henry Joutsijoki
Ulla-Mari Kinnunen
Ann-Kristin Kock-Schoppenhauer
Antoine Lamer
Pashalina Lialiou
Pia Liljamo
Silvia Llorente
Matthias Löbe
Diana Lungeanu
Nestor Adolfo
Romaric Marcilly
Maurice Mars
Sushil Kumar Meher
Carlos Molina
Eustache Muteba Ayumba
Jan Muzik
Pantelis Natsiavas
Andrej Orel
Louise Pape-Haugaard
Carlos Parra
Laura-Maria Peltonen
Giuseppe Rauch
Alejandro Rodríguez González
Philip Scott
Nicola (Nikki) Shaw
Michael Shifrin
Berglind Smaradottir
Martin Staemmler
Milton Stern
Antonis Stylianides
Oscar Tamburis
Tomas Trpisovsky
Philipp Urbauer
Irina Vasilyeva
Jan Vejvalka
Patrick Weber

This page intentionally left blank

Contents

Section I. Artificial Intelligence in Health FAIR Data Driven & FAIR on Infrastructures and Software

Section II. Methods for the Adoption of FAIR Principles & Privacy and Security Aspects Applying FAIR in Health Research

Section III. FAIR and COVID-19 (and Other Infective Diseases) Research Data

Section IV. Metadata, Ontologies and Terminologies to Support Sharing of Health Research Data

Section V. Paradigms to Share Health Research Data & Various Health Informatics Studies

Section I

Artificial Intelligence in Health FAIR Data Driven & FAIR on Infrastructures and Software

This page intentionally left blank

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210799

Federated Mining of Interesting Association Rules Over EHRs

Carlos MOLINA[a,1], Belen PRADOS-SUAREZ[a] and Beatriz MARTINEZ-SANCHEZ[b]
[a] *Software Engineering Department, University of Granada, Spain*
[b] *Computer Science Department, San Cecilio Hospital, Granada, Spain*

Abstract. Federated learning has a great potential to create solutions working over different sources without data transfer. However current federated methods are not explainable nor auditable. In this paper we propose a Federated data mining method to discover association rules. More accurately, we define what we consider as interesting itemsets and propose an algorithm to obtain them. This approach facilitates the interoperability and reusability, and it is based on the accessibility to data. These properties are quite aligned with the FAIR principles.

Keywords. Electronic Health Records, Data Mining, Privacy, Federated Learning

1. Introduction

Nowadays one of the main issues to achieve a proper health data access are the security and the privacy protection. Federated learning [1] has arisen as a solution to deal with them. In this approach, the data sources involved collaborate to learn a model and share what has been learnt with no need for data transfer. To this purpose they use to distribute the calculation between the sources, working locally over the data, and sharing only the calculated values. With no data transfer, the security and privacy protection can be easily achieved.

Most these methods are based on the optimization of numerical values (e.g. neural networks weights by means of Federated Average [2], or Support Vector Machines planes [3]). In those approaches, normally a local model for each data source is learnt. These models are then combined into a global model. However, these approaches have problems when the data distribution is not uniform between the sources, or when it is necessary to adapt the data distributions (see [1] for more details). Moreover, although they obtain good results, is not possible for the user to understand the underlying mathematical model and why a concrete answer is given.

Explainable Artificial Intelligence [4] methods could solve it, and here is where our proposal lays. To our best knowledge there are no federated proposals of data mining techniques as widely used as the Association Rules [5]. With these techniques the answers can be understood by the user, and it is even possible to audit the results to knowthe reasoning inside the learnt models. Our aim here is to find out which rules have interest, as much if they correspond to frequent cases (that affect a great part of the population) as if they are related to infrequent ones (like rare diseases).

[1] Corresponding Author, Carlos Molina, University of Granada; E-mail: carlosmo@ugr.es. ORCID: Carlos Molina (0000-0002-7281-3065), Bele´n Prados-Suarez (0000-0002-3980-102X).

Having federated methods means a step forward in the interoperability since any of the sources can benefit from the results calculated collaboratively. Even more, having a scheme to apply a similar method over different data structures in different sources supports the reusability. Finally, the aggregated response of the coordinator method offers a homogeneous access to all the underlying data sources, which is an improvement in the accessibility to the information, preserving the privacy. It all makes our proposal quite aligned with the FAIR principles [6].

2. Methods

In this section we first present the parameters needed in the process. Next, the algorithm to extract the association rules to work with the EHR data is explained.

The concept of interesting itemset is different from frequent itemset [7]. The latter one represents those facts that occur together in a great number of the cases studied. However, it doesn't allow to discover itemsets that are highly related that but are not quite numerous, like what happens with rare diseases. The purpose of the interesting itemsets is to model not only what is relevant for being frequent, but also what is relevant for being highly related although its global frequency is low. To model this concept, we define the following measure:

Definition 1 *The* Interest *of the itemset* $\{i_1, \dots, i_n\}$ *is the function* In *defined as follows:*

$$In([i_1, \dots, i_n]) = \frac{\dfrac{\sup\,([i_1, \dots, i_n])}{\min\,\{\sup(i_1), \dots, \sup(i_n)\}}}{\max\,\{\sup(i_1), \dots, \sup(i_n)\}} \in [0, +\infty]$$

The function *In* measures how the relative frequency of the items increases when they appear together. For example, a value $In(its) = 2$ means that the relative frequency of the items doubles when they appear together. In our approach, to consider an itemset *its* it has to be frequent ($sup(its) >= threshold_{sup}$) and interesting ($In(its) >= threshold_{In}$). Normally the *consistency* is used to measure the quality of the association rules [5]; but this measure has problems with very frequent items (see [8] for details). To avoid this issue, we follow the proposal of [8,9] to use the *Certainty factor CF* (see [10] for details). This measure was first proposed for an expert system in medicine. In the case of the association rules, the *CF* avoids the *consistency* problems with very frequent itemsets. We consider an association rule *r* when $CF(r) >= threshold_{CF}$.

Once we have presented the measure to be used in the association rules extraction, we present the federated algorithms to work with EHR data. In our approach, we have two different processes, one for the *coordinator* and another for each *node*.

The *Coordinator* algorithm controls the extraction process (Algorithm 1). First (lines 2-4), the *Coordinator* asks the *nodes* to extract the frequent itemset of size 1. When all the nodes have answered, the coordinator checks if there is an identified frequent itemset which does not appear in the answer of some of the nodes (lines 6-11). In that case, the node without that frequent itemset is asked to give its support, so the global support is correctly calculated (function *supportItemset* in Algorithm 2). Let us note that only the coordinator has the global support of the itemsets, so none of the nodes gets information

Table 1. Results of the experiments (Fi=number of interesting itemsets, AR=number of association rules)

Data dist.	Uniform								Random							
No. of nodes	**1**		**2**		**3**		**4**		**1**		**2**		**3**		**4**	
Config.	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**	**F_i**	**AR**
1	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0
2	68	107	68	107	68	107	68	107	68	107	68	107	68	107	68	107
3	290	497	290	497	290	497	290	497	290	497	290	497	290	497	290	497
4	243	436	243	436	243	436	243	436	243	436	243	436	243	436	243	436

from the others nodes in the process. In this step, the *In* function is not applied (all the itemsets have only one item so *In* always values 1). This schema is repeated for itemsets of size 2 and more until for some size we get no interesting itemsets (lines 14-36). The main difference is this case is that we can identify the *Interesting itemsets* using the *In* function to reduce the number of itemsets. In lines 26-30 the coordinator calculates the *In* value of the identified frequent itemsets. An itemset *its* is valid if $In(its) >= threshold_{In}$. After this process, the coordinator sends to each node the interesting itemsets to be considered for next step (function *setInterestingItem- set* in Algorithm 2). The nodes, when generating candidates for frequent itemsets of size *N*, only consider the itemsets that include an interesting itemset of size $N-1$ (line 6 in Algorithm 2). To avoid information transfer, the coordinator only sends to each node the interesting itemsets that have been identified by that node as frequent. With the frequent itemsets calculated, the association rules are generated considering only the rules with a $CF >= threshold_{CF}$ (lines 37-42).

3. Results

To test the proposed algorithm, we used data from COVID-19 patients [11]. In this dataset we have information from 2547 patients. We have tested the methods splitting the data from 1 node (only one node) to 4 nodes, considering uniform distribution and random. In Table 1 we show the results (number of interesting itemsets and association rules) for each of the configurations considering three sets of parameters:

- $Conf_1$=$threshold_{sup} = 0.1, threshpold_{In} = 2, threshold_{CF} = 0.5$;
- $Conf_2$=$threshold_{sup} = 0.05, threshpold_{In} = 5, threshold_{CF} = 0.7$;
- $Conf_3$=$threshold_{sup} = 0.025, threshpold_{In} = 10, threshold_{CF} = 0.7$;
- $Conf_4$= $threshold_{sup} = 0.01, threshpold_{In} = 20, threshold_{CF} = 0.7$.

4. Discussion

The experiments show that the distribution of the data and the number of nodes has no influence on the results. This means that the federated learning process works well independently from the number of nodes and the data distribution.

The proposed method used a synchronous schema of communication. It means that the Coordinator waits for all the nodes to finish each operation. If data distribution is very unbalanced (e.g. one of the data sources has a really greater amount of data than the others) then the *Coordinator* will wait for that *node* to finish meanwhile the other *nodes* and the *Coordinator* are idle. If one of the *nodes* is very slow, due to computational resources or high workload, the *Coordinator* and other *nodes* will have a similar behaviour (waiting for the slow node to finish its operations).

5. Conclusions

We have presented the need for explainable federated mining methods and we have proposed a federated association rule mining algorithm that works with EHR data. It is able to deal with different number of sources and data distributions without quality loose. We have also defined a measure of the interest of an itemset. These federated techniques require a framework that integrates them to take advantage of their potential. We plan to integrate the proposed methods with the EHRagg, [12]. As we have mentioned in the previous section, the synchronous schema has some problems, so an asynchronous proposal that can build an incremental solution would also be interesting.

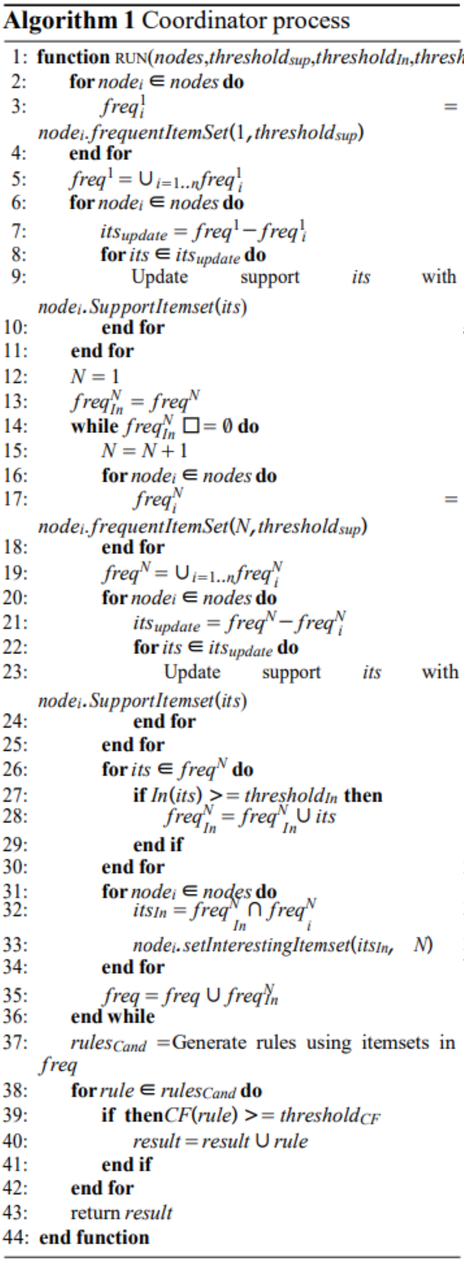
Acknowledgements

This research is partially supported by PGC2018-096156-B-I00 Recuperacio´n y Descrip- cion de Ima´genes mediante Lenguaje Natural usando te´cnicas de Aprendizaje Profundo y Computacio´n Flexible of the Ministerio de Ciencia, Innovacio´n.

References

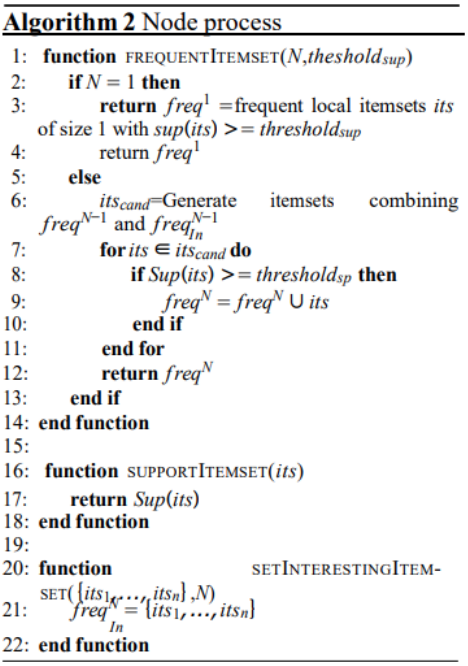
[1] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research. 2021 Mar;5(1):1-9.
[2] McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. InArtificial intelligence and statistics 2017 Apr 10 (pp. 1273-1282). PMLR.
[3] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform. 2018 Apr;112:59-67.
[4] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science Robotics. 2019 Dec 18;4(37).
[5] Srikant R, Agrawal R. Mining generalized association rules.
[6] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9.
[7] Hu J, Li XY. Association rules mining including weak-support modes using novel measures. WSEAS Transactions on Computers. 2009 Mar 1;8(3):559-68.
[8] Delgado M, Marín N, Sánchez D, Vila MA. Fuzzy association rules: general model and applications. IEEE transactions on Fuzzy Systems. 2003 Apr 8;11(2):214-25.
[9] Marin N, Molina C, Serrano JM, Vila MA. A complexity guided algorithm for association rule extraction on fuzzy datacubes. IEEE Transactions on Fuzzy Systems. 2008 Jun 6;16(3):693-714.
[10] Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Mathematical biosciences. 1975 Apr 1;23(3-4):351-79.
[11] MH HOSPITALES. Covid Data Save Lives. https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version, 2021.
[12] Prados-Suárez B, Fernández CM, Yañez CP. Electronic health records aggregators (EHRagg). Methods of Information in Medicine. 2020 May;59(02/03):096-103.

Appendix: Algorithms



Algorithm 1 Coordinator process

```
function RUN($nodes$,$threshold_{sup}$,$threshold_{In}$,$threshold_{CF}$)
    for $node_i \in nodes$ do
        $freq_i^1 = node_i.frequentItemSet(1, threshold_{sup})$
    end for
    $freq^1 = \cup_{i=1..n} freq_i^1$
    for $node_i \in nodes$ do
        $its_{update} = freq^1 - freq_i^1$
        for $its \in its_{update}$ do
            Update support $its$ with $node_i.SupportItemset(its)$
        end for
    end for
    $N = 1$
    $freq_{In}^N = freq^N$
    while $freq_{In}^N \square = \emptyset$ do
        $N = N + 1$
        for $node_i \in nodes$ do
            $freq_i^N = node_i.frequentItemSet(N, threshold_{sup})$
        end for
        $freq^N = \cup_{i=1..n} freq_i^N$
        for $node_i \in nodes$ do
            $its_{update} = freq^N - freq_i^N$
            for $its \in its_{update}$ do
                Update support $its$ with $node_i.SupportItemset(its)$
            end for
        end for
        for $its \in freq^N$ do
            if $In(its) >= threshold_{In}$ then
                $freq_{In}^N = freq_{In}^N \cup its$
            end if
        end for
        for $node_i \in nodes$ do
            $its_{In} = freq_{In}^N \cap freq_i^N$
            $node_i.setInterestingItemset(its_{In}, N)$
        end for
        $freq = freq \cup freq_{In}^N$
    end while
    $rules_{Cand}$ =Generate rules using itemsets in $freq$
    for $rule \in rules_{Cand}$ do
        if then$CF(rule) >= threshold_{CF}$
            $result = result \cup rule$
        end if
    end for
    return $result$
end function
```



Algorithm 2 Node process

```
function FREQUENTITEMSET($N$,$theshold_{sup}$)
    if $N = 1$ then
        return $freq^1$ =frequent local itemsets $its$ of size 1 with $sup(its) >= threshold_{sup}$
        return $freq^1$
    else
        $its_{cand}$=Generate itemsets combining $freq^{N-1}$ and $freq_{In}^{N-1}$
        for $its \in its_{cand}$ do
            if $Sup(its) >= threshold_{sp}$ then
                $freq^N = freq^N \cup its$
            end if
        end for
        return $freq^N$
    end if
end function

function SUPPORTITEMSET($its$)
    return $Sup(its)$
end function

function SETINTERESTINGITEMSET($\{its_1, \ldots, its_n\}$,$N$)
    $freq_{In}^N = \{its_1, \ldots, its_n\}$
end function
```

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210800

Encoding Health Records into Pathway Representations for Deep Learning

Marco Luca SBODIO[a,1], Natasha MULLIGAN[a], Stefanie SPEICHERT[a],
Vanessa LOPEZ[a] and Joao BETTENCOURT-SILVA[a]
[a] *IBM Research Europe*

Abstract. There is a growing trend in building deep learning patient representations from health records to obtain a comprehensive view of a patient's data for machine learning tasks. This paper proposes a reproducible approach to generate patient pathways from health records and to transform them into a machine-processable image-like structure useful for deep learning tasks. Based on this approach, we generated over a million pathways from FAIR synthetic health records and used them to train a convolutional neural network. Our initial experiments show the accuracy of the CNN on a prediction task is comparable or better than other autoencoders trained on the same data, while requiring significantly less computational resources for training. We also assess the impact of the size of the training dataset on autoencoders performances. The source code for generating pathways from health records is provided as open source.

Keywords. Patient Representation, Convolutional Neural Networks, EHR

1. Introduction

Despite the promising results of deep learning techniques for performing analytics tasks, several open challenges remain in dealing with the heterogeneous data from Electronic Health Records (EHRs) and the lack of model intelligibility and interpretability required for real-world applications [1],[2],[3]. Data representation and encoding plays a key role in training successful models for prediction tasks and explainability; additionally, compact representations have been used to address challenges such as sparseness of EHR data [4]. Multi-source EHR data has been modelled as patient trajectories through time [5] or by representing EHR information as a 2D matrix [6] of appointments and diagnoses codes where convolutional neural networks (CNNs) were used for risk prediction [7]. The most common type of representation is a sequential ordering of a patient's data used as input to a Recurrent Neural Network (RNN) for application areas such as, for example, prediction or phenotyping [8].

This paper proposes a pathway representation that maps patient's health records into different classes over time, to form a machine-processable image-like structure for further analyses and deep learning tasks. A Patient Pathway Extractor application was developed and used to transform EHR data into the new representations (described in section 2.2). The application is shared in an open-source git repository. The pathways generated using the proposed representations were then validated using three mainstream

[1] Corresponding Author, IBM Research Europe, Dublin, Ireland; E-mail: marco.sbodio@ie.ibm.com

deep learning algorithms, described in section 2.3, and the preliminary results using synthetic data are included in section 3.

2. Methods

2.1. Pathway Representation

We propose an encoding of pathway data with an accompanying open-source application called Patient Pathway Extractor[2]. We use a set of predefined classes to classify data and pathway events; we build a structured representation that shows the discretized values of the data along a time dimension. We use Synthea [7], a FAIR dataset generator, to produce the CSV input for the Patient Pathway Extractor. More precisely, we classify data generated by Synthea along the following classes: *demographics* (patient details), *observations* (results of clinical exams and vitals), *conditions* (diagnoses and care plans), *medications*, *procedures* and *outcomes* (readmission, death, survival at a point in time).

Data may consist of isolated events happening at a specific point in time, of events having a duration. Events can be visualized along a timeline: isolated events can be shown as dots, while events having a duration can be shown as horizontal bars. When multiple events happen at the same time, or when an event include a set of data values, the timeline visualization can display these using an overlay information box.

Humans can easily understand the timeline visualization of a pathway, but such a representation is not helpful when trying to analyse data using machine learning or deep learning algorithms. For this reason we propose a novel image-like representation of the pathway data. We build such image-like representation using a three steps process: (1) representing the discretized data points in a 3-dimensional grid, (2) projecting into a bi-dimensional grid, and (3) numerical encoding.

Firstly, we map the discretized values of the data in a 3-dimensional grid. The dimensions of the grid represent respectively the order of the events (time), the different classes (demographics, observations, conditions, medications, procedures and outcomes), and co-occurrence of events (values of a given class having the same timestamp). Figure 1 shows (on the left) the 3-dimensional grid representation of the pathway timeline. Note that we do not encode timestamps along the time dimension, but only retain the order of events (recording timestamps is possible with a simple extension of the proposed representation). We use a configurable set of rules that discretize values into custom bins. We use spreadsheet (easily interpretable by practitioners) to define the rules, and parse them into executable formats using the Drools[3] rule engine. Our current sets of 246 rules cover demographics, medications, observations (based on patient age and gender, the LOINC code and its units), as well as outcomes. For example, a rule that takes as input a body mass index (BMI) observation (LOINC code 39156-5) and when its value is in the range [18.5, 25 kg/m^2], it maps it to the bin value “normal BMI”.

Subsequently, we project the 3-dimensional grid into a bi-dimensional grid: the horizontal axis denote the order of events, while the vertical axis denote the various classes of our representation. The projection places values having the same timestamp (co-occurring) one after the other along the horizontal axis. The order of events is

[2] Pathway Extractor, https://github.com/Alvearie/patient_pathway_extractor/

[3] Drools, https://www.drools.org

preserved along the x-axis. The pathway representation does not impose any restriction on the order of the classes on the y-axis and downstream applications may use different ordering. In practice, we consider every bi-dimensional corresponding to a value along the time dimension, we rotate each slice along the class dimension, and finally concatenate them as illustrated in Figure 1.

As a final step, we use a numeric space to encode the values of the bi-dimensional grid in a numeric space. The encoding space is ℝ but can also be ℕ depending on the downstream analysis task; for some applications we may encode values in the RGB space, which translates our representation into an image. This encoding step of the process produces a numeric representation of the pathway, which, while retaining meaningful dimensions, is also easy to use as input for machine learning and deep learning tasks.

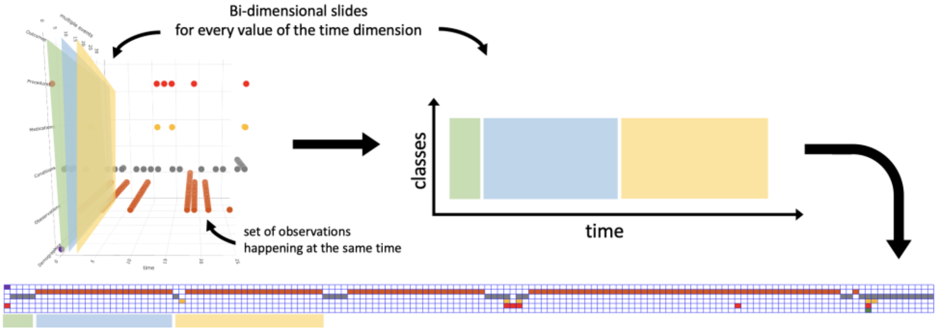

Bi-dimensional slides
for every value of the time dimension
set of observations
happening at the same time
classes
time

Figure 1. Transformation of patient pathway data from the 3-dimensional grid to the final bi-dimensional grid representation.

2.2. Synthetic EHR Data and Pathways Generation

We tested our approach using synthetic data generated by Synthea [9]. The generator creates realistic patient data with the help of a rule-based backend that determines the course of a condition. As with real-world EHR data, patients may have multiple concurrent conditions, which, over the course of their life, may interact or influence each other. Using Synthea, we generated a population of 500,000 patients, from which we extracted 1,073,105 pathways considering only a set of ten conditions including common chronic (e.g. diabetes) and acute (e.g. appendicitis, fractures) conditions.

2.3. Deep Learning Frameworks used to test the Pathway Representation

We used our pathways representation to create three pathway encodings using three types of autoencoders: Multilayer perceptron (MLP) autoencoder (Denoising Autoencoder), Sequence-to-Sequence (RNN) autoencoder, and CNN autoencoder.

An analysis of the pathways in our dataset has revealed that the maximum pathway length was 5128 data points on the x-axis; however, around 98% of the pathways fitted into a 6*400 grid (6 classes by 400 points on the x-axis). Pathways with a length of less than 400 were padded with zeros and only pathways not exceeding this size were used for training and testing with an 80/20% split. All autoencoders were designed to generate pathway encodings of the same length and their architectures are described in Figure 2.

- The MLP autoencoder consists of 3 layers for both encoder and decoder and a vector (bag) of pathway events was used as an input. Temporal dimension is not supported.

- The RNN autoencoder was built using GRU cells. The pathway grid was sliced vertically (time axis) and all events in that slice were concatenated providing a sequence of input vectors.
- The CNN autoencoder directly supports the pathways as input. It has an encoder with three convolutional layers followed by a fully connected layer to produce the pathway encoding. The decoder has the same architecture but in the reverse order.

We evaluated the performance of the three autoencoders using the following prediction task: given an input pathway, we remove from its representation the data identifying the medical condition that originated the patient pathway, and we use the trained autoencoder to predict such condition.

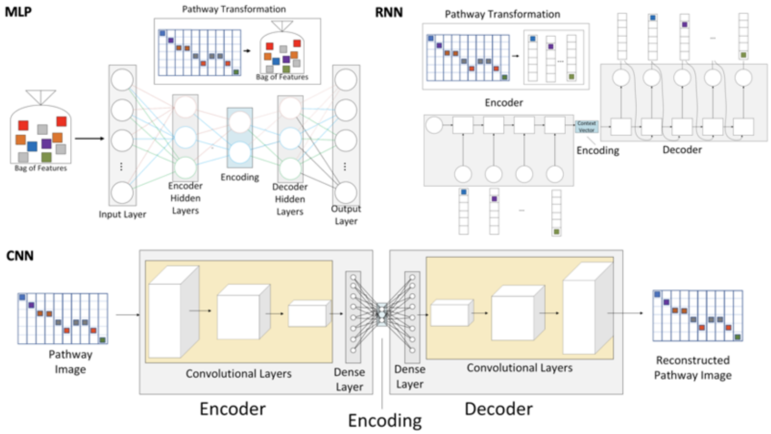
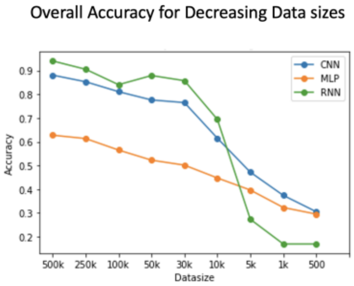


Figure 2. The three architectures used for compact representation of pathways.

3. Results and Discussion

In our experiments, the Pathway Extractor generated 1,073,105 pathways from a large EHR dataset. The flexible way in which pathways and conditions are represented allows for new classes to be added or existing ones to be dropped. Similarly, the codes and values may be adapted by modifying the discretization rules. Our approach is not limited to Synthea and may be applied to other EHR datasets.

In our prediction task, RNN gave the best accuracy (94.0%) followed by CNN (88.1%) and MLP (62.8%). We then tested the models performance using different training datasets to understand the impact on their accuracy. Figure 3 shows the overall accuracies for the three autoencoders when trained on the pathway data extracted from synthetic populations (generated with Synthea) of decreasing size (from 500,000 to 500 people). CNN and RNN, as expected, outperformed MLP, and larger training sets increase prediction accuracies. We note that for RNN and CNN there is a significant increase in accuracy with training sets computed with a population larger than 10,000. Overall CNN achieves good accuracy while requiring considerably less computational resources for training compared to RNN: 37.5% less memory, and over 98% less time (see Figure 3). Additionally, CNN and our pathway image-like representation may help in explaining predictions by using existing techniques such as attention to highlight important events in the input pathway grid [5].


Overall Accuracy for Decreasing Data sizes

Accuracy and computational resources for the population size of 500,000

	CNN	RNN	MLP
Accuracy (%)	88.1	94.0	62.8
Time (min)	15	1320	65
RAM (GB)	25	40	25

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Figure 3. Accuracies and computational resources of the three autoencoders.

4. Conclusions

This paper describes an approach to represent patient pathways and provides an accompanying open-source application that transforms health records into a machine-processable representation for deep learning tasks. We have evaluated this approach using data generated by Synthea, and observed that in a prediction task a CNN performed almost as well as an RNN, while being significantly less expensive to train in terms of computational resources and training time, and enabling further work on predictions explainability. Our results also give insight on how much data may be required for model training. Further work includes expanding the pathway representation with additional classes and data beyond EHRs.

References

[1] Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, Jim Zheng W, Roberts K. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. J Biomed Inform. 2021 Mar;115:103671..

[2] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018 Sep;22(5):1589-1604. doi: 10.1109/JBHI.2017.2767063. Epub 2017 Oct 27.

[3] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016 May 17;6:26094.

[4] Yuqi S, Jingcheng D, Zhao L, Xiaoqian J, Timothy M, Fe, W, Zheng WJ, Kirk R. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. Journal of Biomedical Informatics 115 (2021), pp.103671

[5] Nguyen-Duc T., et al. Deep EHR Spotlight: a Framework and Mechanism to Highlight Events in Electronic Health Records for Explainable Predictions. AMIA 2021 Virtual Informatics Summit (2021)

[6] Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: A deep learning approach. InProceedings of the 2016 SIAM International Conference on Data Mining 2016 Jun 30 (pp. 432-440). Society for Industrial and Applied Mathematics.

[7] Suo Q, Ma F, Yuan Y, Huai M, Zhong W, Gao J, Zhang A. Deep patient similarity learning for personalized healthcare. IEEE transactions on nanobioscience. 2018 May 16;17(3):219-27.

[8] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. InMachine learning for healthcare conference 2016 Dec 10 (pp. 301-318). PMLR.

[9] 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. 2018 Mar 1;25(3):230-8.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210801

A Learning Framework for Medical Image-Based Intelligent Diagnosis from Imbalanced Datasets

Tetiana BILOBORODOVA[a,1], Inna SKARGA-BANDUROVA[b], Mark KOVERHA[c],
Illia SKARHA-BANDUROV[d] and Yelyzaveta YEVSIEIEVA[e]
[a]*G.E. Pukhov Institute for Modelling in Energy Engineering, Ukraine*
[b]*Oxford Brookes University, United Kingdom*
[c]*Volodymyr Dahl East Ukrainian National University, Ukraine*
[d]*Luhansk State Medical University, Ukraine*
[e]*School of Medicine, V. N. Karazin Kharkiv National University, Ukraine*

Abstract. Medical image classification and diagnosis based on machine learning has made significant achievements and gradually penetrated the healthcare industry. However, medical data characteristics such as relatively small datasets for rare diseases or imbalance in class distribution for rare conditions significantly restrains their adoption and reuse. Imbalanced datasets lead to difficulties in learning and obtaining accurate predictive models. This paper follows the FAIR paradigm and proposes a technique for the alignment of class distribution, which enables improving image classification performance in imbalanced data and ensuring data reuse. The experiments on the acne disease dataset support that the proposed framework outperforms the baselines and enable to achieve up to 5% improvement in image classification.

Keywords. Medical image classification, imbalanced data, machine learning oversampling

1. Introduction

The recent success of machine learning (ML) and computer vision allows elevating medical diagnostics to a new level, particularly in the classification of visual data enabling to solve problems of medical analytics and clinical decision making more rapid and accurate. Being a key component of intelligent diagnosis, medical images classification includes identifying the features in an image and predicting the class of a specific object in an image. In this context, data quality has a significant impact on the success of ML algorithms and is a core of scientific knowledge. As mentioned in [1], the ideal medical image dataset for the ML application is Findable, Accessible, Interoperable, and Reusable (FAIR) [2]. This basically means that image datasets should have adequate volume and class distribution, be well-annotated, verifiable, ground-truth, and reusable. However, in the wild, due to their nature, medical image datasets are being the long way

[1] Corresponding Author, Tetiana Biloborodova, G.E. Pukhov Institute for Modelling in Energy Engineering, Kyiv, Ukraine; E–mail: beloborodova.t@gmail.com.

of the FAIR principles, and in many cases, they are closed, limited distributed, relatively few annotated and highly imbalanced.

Meanwhile, the ML algorithms require a large number of annotations, which is a time-consuming and labour-intensive process and often, the class distribution inside datasets is not equal. This is because of identifying and predicting process often includes rare events [3]. Medical data demonstrate an uneven distribution of classes in rare clinical cases or diseases, which makes it difficult to form a balanced dataset for training which in turn leads to poor reproducibility of the ML algorithms. Rare cases result in data imbalance, namely the imbalance in the number of objects in different classes. Imbalanced data refers to a dataset where the class distribution is not uniform among the classes. The prevailing class is called the majority class, and the smallest class in terms of objects is the minority class [4]. Imbalanced data can negatively affect the accuracy of the models and lead to incorrect or erroneous classification results.

To following the FAIR principles, this study proposes the technique for dealing with heavily imbalanced datasets and introduces the concept of a machine-readable data preprocessing and resampling for model learning. We aim to extend previous research in imbalanced data classification, improve quality of computer vision-based disease diagnostic and provide usability and reusability of medical image datasets. Inoculation of the FAIR principles as a new data management strategy results in significant improvements in automation of medical image diagnostic through machine readability and enable reuse data and improve their scalability.

2. Methods

Methods to handle imbalanced data can be divided into three large categories: data-layer methods, algorithm-layer methods, and cost-sensitive learning methods [5]. Data layer methods include resampling (oversampling, undersampling and hybrid) techniques. This is the most straightforward and widely adopted approach for dealing with highly imbalanced datasets. All of these techniques follow FAIR principles in part of being machine-readable for reusable. Algorithm-level methods include the use of essemble methods based on machine learning algorithms [6]. Cost-sensitive learning methods target the problem of imbalanced learning by using other evaluation metrics and different cost matrices that describe the costs for misclassifying any particular data example [7].

2.1. Imbalanced datasets

We define imbalanced dataset S with m objects, $|S| = m$, as $S = \{(x_i, y_i), i = 1, ..., m$, where $x_i \in X$ is an object in n-dimension space of input features $X=\{f_1, f_2, ..., f_n\}$, and $y_i \in Y = \{1, …, C\}$is the label, associated with object x_i. At it simplest, $C = 2$ means the binary classification task where two subsets are defined as $S_{min} \subset S$ the minor class subset S_{min} in S and $S_{maj} \subset S$ is a major class subset such as $S_{min} \cap S_{maj} = \{\Phi\}$ and $S_{min} \cup S_{maj} = \{S\}$.

The objects generated from the dataset S are defined as E, with disjoint subsets E_{min} and E_{maj} that represent the minority and majority classes of E, respectively, each time they are used.

2.2. Proposed approach

The basic structure of proposed approach to obtaining an accurate model for classification of medical images in the imbalanced datasets is shown in Fig. 1.

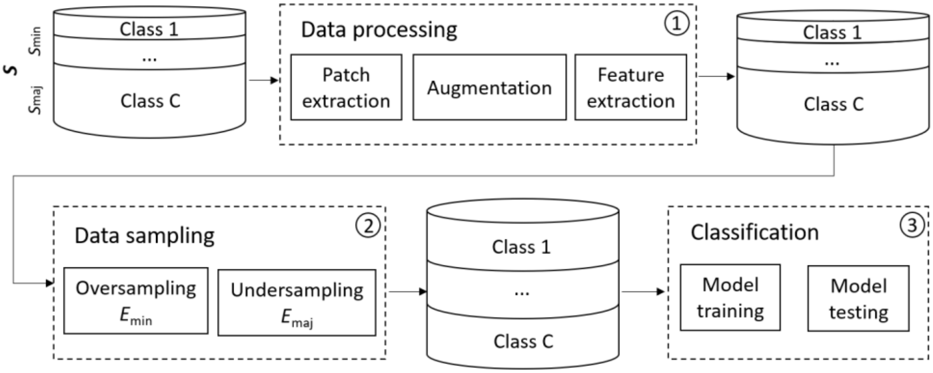


Figure 1. The structure of proposed approach

The learning framework for medical image-based diagnosis from imbalanced datasets incorporates data processing, data sampling, and classification. Since we are dealing with imbalanced datasets, collected data are being annotated and analysed in terms of minority and majority classes. The data processing phase includes automatic patch extraction, data augmentation and feature extraction. The output data at this phase are the extracted features. For phase 2, class distribution is evaluated, and the resampling technique is selected depending on the size and type (minority or majority) of the imbalance in different classes. It can be done by removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling). Oversampling is used for sampling minority class objects, while undersampling is used for sampling majority class objects. From this phase, we obtain a quasi balanced machine-readable dataset ready for model training. Finally, the model training and validation are performed.

3. Results

To evaluate the performance of the proposed technique, the experiments with an open medical image dataset ACNE04 provided by Wu et al. [8] were conducted. The dataset includes 1457 face images and expert annotations according to the Japanese acne grading scale. There are four acne severity classes, namely 0 Mild equal to 410 samples, 1 Moderate equal to 506 samples, 2 Severe equal to 146 samples, and 3 Very severe equal to 103 samples. In general, we used 1165 images to train the model and 291 images to test the model, which corresponds to a distribution of 80% for training and 20% for testing. Proposed approach runs on an NVIDIA GeForce GTX 1060 with 3 GB VRAM and is implemented based on the PyTorch framework.

The data processing phase included patch extraction, data augmentation and feature extraction. At the patch extraction stage, we utilised two pre-trained models: (1) shape_predictor_68_face_landmarks model [9] and (2) the One Eye model [10]. In case when any of these models could not process the image, the entire original image was used for the next phase. Each patch inherited the label of the original image and had a binding to it, which was used later to get a general estimate of the degree severity of acne for a photo. For augmentation, a sliding translation of patches was used. Further, the feature extraction for each patch was carried out using the ResNet-152 model [11]. Data distribution by classes after augmentation is presented as follow. 0 Mild: 3556 samples, 1 Moderate: 4333 samples, 2 Severe: 1843 samples, and 3 Very severe: 1514 samples. Each patch is bound to the original image, and thus the extracted features inherit the dependencies of the patches. Then, the classes were revised following acne severity grades, and extracted features were used for sampling. Sampling and minority class generation were done via Synthetic Minority Oversampling Technique (SMOTE) [12]. The total number of samples for each class was fitted to the most numerous class after feature extraction and equal to 4333 for each class.

Data generated at the oversampling phase was used to train a convolutional neural network (CNN) model and estimate the severity of the acne from the face image. Model training was run for 17985.2 seconds. The classification problem was transformed into a regression task at this phase by defining acne severity grades as integer equivalents. It was done to reduce possible subjectivity in the expert's annotation of the acne severity. The inverse transformation was done using [0.5, 1.5, 2.5] as the edge list. Model evaluation using the trained CNN is implemented on test data. Since the problem was reduced to a regression problem, the corresponding criteria for assessing the regression quality were used. As a result, the following values were obtained for the ACNE04: RMSE = 0.397419, EV = 0.826736, MAPE = 0.199264, R2 = 0.826682.

4. Discussion

In order to test effectiveness of the proposed approach, we compared our results with outcomes without oversampling (Table 1). Classification accuracy was calculated after converting the results back from continuous to discrete scale using [0.5, 1.5, 2.5] as a list of edges. Delta was calculated as the difference in the results between these two approaches in percentage.

Table 1. The results of experiments with the basic and proposed approach

Metrics	Without oversampling	Proposed approach	Delta, %
RMSE	0.422356	0.397419	5.904261
EV	0.826736	0.873874	5.7017
MAPE	0.199264	0.171855	13,75512
R^2	0.826682	0.873646	5.68102
Accuracy	80 %	85 %	5

Both RMSE and MAPE show a smaller error, while the EV and R^2 criteria show higher values for proposed approach, which indicates a higher quality of the model. Comparison of the obtained results with two benchmark models is presented in Table 2.

Table 2. Comparison of acne classification research

Approach for acne classification	Accuracy (%)	ER (%)
Wu et al. [8]	84.11	15.89
Lim et al. [13]	67	33
Ours	85	15

As can be seen from the table, the proposed approach showed the highest accuracy and the lowest error rate in comparison with studies with imbalanced data, however, it should be mentioned that it did not outperform the results for study [14] where data was initially balanced (stated accuracy 99.44%), which sounds natural but needs further investigation.

5. Conclusion

Accurate classification of medical images is one of the first steps towards the wide adoption of computer vision into healthcare industry. In this paper, we propose a complex approach for imbalanced medical image classification. It is grounded on FAIR prinsiple where medical image datasets remain useful even in high inbalance and can be reused to train, test, validate, verify, and regulate ML products. Experiments on acne image dataset showed that the proposed learning framework able to improve the classification performance metrics and proved their advantages in comparison with basic approach without oversampling.

References

[1] Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. Journal of digital imaging. 2017 Aug;30(4):392-9.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. Andra Waagmeester. Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016.

[3] Weiss GM, Hirsh H. Learning to predict extremely rare events. InAAAI workshop on learning from imbalanced data sets 2000 Jul (pp. 64-68). Austin: AAAI Press.

[4] Yijing L, Haixiang G, Xiao L, Yanan L, Jinling L. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowledge-Based Systems. 2016 Feb 15;94:88-104.

[5] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications. 2017 May 1;73:220-39..

[6] *Imbalanced learning: Foundations, Algorithms, Applications* (Eds.: Haibo He, Yunqian Ma) Wiley-IEEE Press, 2013.

[7] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 1–48 (2019).

[8] Wu X, Wen N, Liang J, Lai YK, She D, Cheng MM, Yang J. Joint acne image grading and counting via label distribution learning. InProceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 10642-10651)..

[9] Davisking/dlib-models. Github.com. [Online] https://github.com/davisking/dlib-models

[10] OpenCV. Github.com. [Online] https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_eye.xml.

[11] The Microsoft Cognitive Toolkit. Cntk.ai. [Online] https://www.cntk.ai/Models/Caffe_Converted/ResNet152_ImageNet_Caffe.model

[12] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57.

[13] Lim ZV, Akram F, Ngo CP, Winarto AA, Lee WQ, Liang K, Oon HH, Thng ST, Lee HK. Automated grading of acne vulgaris by deep learning with convolutional neural networks. Skin Research and Technology. 2020 Mar;26(2):187-92.

[14] Junayed MS, Jeny AA, Atik ST, Neehal N, Karim A, Azam S, Shanmugam B. Acnenet-a deep cnn based classification approach for acne classes. In2019 12th International Conference on Information & Communication Technology and System (ICTS) 2019 Jul 18 (pp. 203-208). IEEE.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210802

Assessing Acceptance Level of a Hybrid Clinical Decision Support Systems

Georgy KOPANITSA[a,1], Ilia V. DEREVITSKII[a], Daria A. SAVITSKAYA[b] and Sergey V. KOVALCHUK[a,b]
[a] *ITMO University, 49 Kronverskiy prospect, 197101, Saint Petersburg, Russia*
[b] *Almazov National Medical Research Centre, 2 Akkuratova st., 197341, Saint Petersburg, Russia*

Abstract. We present a user acceptance study of a clinical decision support system (CDSS) for Type 2 Diabetes Mellitus (T2DM) risk prediction. We focus on how a combination of data-driven and rule-based models influence the efficiency and acceptance by doctors. To evaluate the perceived usefulness, we randomly generated CDSS output in three different settings: Data-driven (DD) model output; DD model with a presence of known risk scale (FINDRISK); DD model with presence of risk scale and explanation of DD model. For each case, a physician was asked to answer 3 questions: if a doctor agrees with the result, if a doctor understands it, if the result is useful for the practice. We employed a Lankton's model to evaluate the user acceptance of the clinical decision support system. Our analysis has proved that without the presence of scales, a physician trust CDSS blindly. From the answers, we can conclude that interpretability plays an important role in accepting a CDSS.

Keywords. CDSS, user acceptance, data-driven, rule-based

1. Introduction

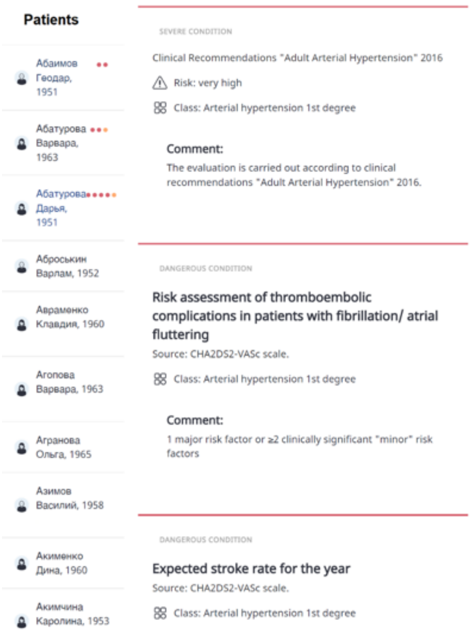
Clinical decision support systems (CDSS) are made to support evidence-based patient care and shared-decision making to improve health and wellbeing of patients. While various studies have demonstrated that CDSSs decrease medical errors and improve clinical outcomes, we can see that CDSSs did not yet reach their full potential due to low acceptance and adoption [1,2]. Among the factors that influence adoption and acceptance we can name relevance of the provided information and the validity of the system [3]. Validity and ability to interpret the decision support output can be especially problematic for data-driven CDSSs [4]. One of the approaches to solve the interpretability problem is a hybrid approach [5] where a data-driven decision support is complimented by rule-based methods and scales. We have implemented a CDSS for diabetes complications management using a three-stage hybrid approach [6]. The goal of this study is to understand how a combination of data-driven decision support methods with a rule-based interpretation affect the acceptance and adoption by doctors.

[1] Corresponding author, Georgy Kopanitsa, ITMO University, Saint-Petersburg, Russia, E-mail: georgy.kopanitsa@gmail.com

2. Method

2.1. Decision support system

The CDSS in focus of this study predicts 5 years risks of type 2 diabetes (T2DM) mellitus complications [7]. It includes machine learning based inference along with a FINDRISK scale [8]. The model that is the basis of the CDSS does not require sophisticated medical tests and provides the following prediction efficiency: sensitivity of 76.0% and specificity of 60.2%. The interface of the CDSS with a synthetic data is shown in the figure 1.



Patients

Абаимов ●● Геодар, 1951

Абатурова ●●● Варвара, 1963

Абатурова●●●●● Дарья, 1951

Аброськин Варлам, 1952

Авраменко Клавдия, 1960

Агопова Варвара, 1963

Агранова Ольга, 1965

Азимов Василий, 1958

Акименко Дина, 1960

Акимчина Каролина, 1953

SEVERE CONDITION

Clinical Recommendations "Adult Arterial Hypertension" 2016

Risk: very high

Class: Arterial hypertension 1st degree

Comment:

The evaluation is carried out according to clinical recommendations "Adult Arterial Hypertension" 2016.

DANGEROUS CONDITION

Risk assessment of thromboembolic complications in patients with fibrillation/ atrial fluttering

Source: CHA2DS2-VASc scale.

Class: Arterial hypertension 1st degree

Comment:

1 major risk factor or ≥2 clinically significant "minor" risk factors

DANGEROUS CONDITION

Expected stroke rate for the year

Source: CHA2DS2-VASc scale.

Class: Arterial hypertension 1st degree

Figure 1. CDSS interface

The structure of the study is based on the theory of planned behavior (TPB). It considers attitude, subjective standards, and perceived behavioral control influencing behavioral intentions (and actual behavior).

2.2. CDSS efficiency

We estimated a perceived usefulness of the systems. This metric represents a degree, to which users suppose that utilizing a decision support system will increase their efficiency. We have conducted a survey with physicians who have experience of operating the system. The survey was structured into two phases. The evaluate the perceived usefulness we randomly generated CDSS output in three different settings:

- (A) Data-driven model output,
- (B) Data-driven model with a presence of FINDRISK scale
- (C) FINDRISK scale and explanation.

For each type of settings, we randomly generated a questionnaire with synthetic T2DM cases and questions. Each case was presented to a physician with a patient's basic information and vital signs including antithrombotic therapy (AH), physical activity, blood sugar, short hereditary anamnesis, blood pressure) and a setting specific (A-C) CDSS output.
For each case, a physician was asked to answer 3 questions:

- if a doctor agrees with the result,
- if a doctor understands it,
- if the result is useful for the practice.

All the questions could be answered using a Likert scale with 5 points from 1 (strongly disagree) to 5 (strongly agree).

2.3. User acceptance

The acceptance of the decision support system was evaluated using a Wilson's model of electronic health solutions' acceptance modified by Lankton [9]. The model enables assessing the following metrics: behavioral intention to use (BI), intrinsic motivation (IM), perceived ease-of-use (PEOU), and perceived usefulness (PU) of the decision support system. We measured BI and PU using 2 objects for each metric. IM and PEOU metrics were measured with 3 objects. To rate each item, we applied Likert scale with 5 points: from 1 (strongly disagree) to 5 (strongly agree):

1. Behavioral intention to use
    a. I will use the CDSS to have a second opinion on the patient's risks
    b. I believe I will utilize the CDSS in my practice
2. Intrinsic motivation
    a. The CDSS helps me to make better informed decisions
    b. I trust the CDSS as it provides interpretations of the output
    c. I trust the system as it provides references to the standard scales
3. Perceived ease of use
    a. The CDSS outcomes are clear and understandable
    b. The interpretations are clear, and I understand the reasoning
    c. The visualizations are well-defined, and I don't spend much time on their interpretation
4. Perceived usefulness
    a. CDSS improves the effectiveness of managing risks of patients
    b. It explains me why a certain risk assessment is done

After we collected and analyzed the results of the user acceptance evaluation, we have organized a study to deeper understand the reasoning of the doctors when working with the CDSS. We designed a study as a series of semi-structured one-to-one interviews with an interview script [10], which was created and approved by the research team.

1. Can you understand a model output without interpretations?
2. Are you convinced with the interpretations that the system provides?
3. Do you require an interpretation to critically assess a model output?
4. Does a reference to a scale facilitate assessment of a recommendation?
5. Can you please give any improvement comments or suggestions?

3. Results

We have gathered 161 answers with equal distribution for each setting: 53, 55, 53 for A, B, C, respectively (Table 1).

Table 1. Case scoring mean (95% confidence interval)

	Agree	Understand	Use
Setting A	4.05 (3.78, 4.32)	4.64 (4.42, 4.85)	3.8 (3.05, 4.80)
Setting B	3.16 (2.94, 3.38)	3.98 (3.66, 4.29)	3.38 (3.11, 3.64)
Setting C	3.41 (3.15, 3.67)	4.24 (3.90, 4.54)	3.52 (3.20, 3.85)
All settings	3.54 (3.38, 3.69)	4.28 (4.12, 4.45)	3.56 (3.39, 3.73)

The median values for behavioral intention to use, intrinsic motivation, perceived ease-of-use, and perceived usefulness (PU) demonstrated a general acceptance of the CDSS by the users (Table 2).

Table 2. CDSS acceptance metrics

Metric, Item	Median	Max	Min
1. Behavioral intention to use	3	5	2
1a. I will use the CDSS to have a second opinion on the patient’s risks	3	5	3
1b. I believe I will utilize the CDSS in my practice	3	4	2
2. Intrinsic motivation	3	4	2
2a. The CDSS helps me to make better informed decisions	3	4	2
2b. I trust the system as it provides interpretations of the results	3	4	3
2c. I trust the system because it provides references to the standard scales	3	4	2
3. Perceived ease of use	4	5	2
3a. The model outcomes are clear and understandable	4	5	3
3b. The interpretations are clear, and I understand the reasoning	4	4	2
3c. The visualizations are clear and I don’t spend much time on their interpretation	4	5	3
4. Perceived usefulness	4	5	2
4a. CDSS improves the effectiveness of managing risks of patients	4	4	2
4b. It explains me why the a certain risk assessment is done	4	5	3

4. Discussion and Conclusions

The analysis showed the highest scores were obtained in Setting A, while the lowest is obtained in Setting B. Our interview analysis has proved that without the presence of a standard scale, a physician trust blindly a CDSS results. This can increase type I errors, which can be lowered in comparison to the basic scales. We have analyzed the answers of the participating doctors to understand the reasoning behind the acceptance evaluation. From the answers of the doctors, we can conclude that interpretability provided by rule-based scales play an important role in understanding and accepting a CDSS output, especially when interpretation is done on the feature basis. The system’s output is convincing, and the doctors can act upon it. Interpretations also help doctors to identify incorrect conclusions when a system produces them. They still see room for improvements, as not everything should be measured in numbers. The doctors see the importance of combining data-driven output with rule-based scales. Despite the understanding that data-driven models are based on high-quality, real-world data, doctors still ask for standard and known tools as they are created from the formal research results and widely accepted clinical guidelines. Doctors still believe that experts should be involved in the model development. This can potentially help to expose the results of the CDSS even to patients. Our results show that a hybrid approach when a data-driven models are complimented with standard rule-based scales increases its acceptance and usefulness.

Acknowledgments

This work financially supported by Ministry of Education and Science of the Russian Federation, Agreement #13.2251.21.0067. Unique Identification RF----225121X0067. The work of Georgy Kopanitsa was financially supported by the Government of the Russian Federation through the ITMO fellowship program.

References

[1] Wilson EV, Lankton N. Effects of Prior Use, Intention, and Habit on IT Continuance Across Sporadic Use and Frequent Use Conditions. Communications of the Association for Information Systems [Internet]. 2013 Sep 1 [cited 2021 Jul 20];33(1):3. Available from: https://aisel.aisnet.org/cais/vol33/iss1/3

[2] Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018. 2018 Jul 24;447.

[3] de Clercq PA, Blom JA, Korsten HHM, Hasman A. Approaches for creating computer-interpretable guidelines that facilitate decision support. Artificial Intelligence in Medicine. 2004 May;31(1):1–27.

[4] E C. Data-driven clinical decision processes: it's time. Journal of translational medicine [Internet]. 2019 Feb 12 [cited 2021 Aug 3];17(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30755218/

[5] Derevitskii I, Funkner A, Metsker O, Kovalchuk S. Graph-Based Predictive Modelling of Chronic Disease Development: Type 2 DM Case Study. Studies in health technology and informatics. 2019;261:150–5.

[6] Kovalchuk SV, Kopanitsa GD, Derevitskii IV, Savitskaya DA. Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. 2020 Jul 25 [cited 2021 Aug 6]; Available from: https://arxiv.org/abs/2007.12870

[7] Elkhovskaya L, Kabyshev M, Funkner A, Balakhontceva M, Fonin V, Kovalchuk S. Personalized Assistance for Patients with Chronic Diseases Through Multi-Level Distributed Healthcare Process Assessment. Studies in health technology and informatics. 2019;261:309–12.

[8] Kahl KG, Schweiger U, Correll C, Müller C, Busch ML, Bauer M, Schwarz P. Depression, anxiety disorders, and metabolic syndrome in a population at risk for type 2 diabetes mellitus. Brain Behav. 2015 Mar;5(3):e00306

[9] Lankton NK, Wilson EV, Mao E. Antecedents and determinants of information technology habit. Information and Management. 2010;47(5–6).

[10] Ahlin E. Semi-Structured Interviews With Expert Practitioners: Their Validity and Significant Contribution to Translational Research. Semi-Structured Interviews With Expert Practitioners: Their Validity and Significant Contribution to Translational Research. 2019 Jan 3;

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210803

A Data-Driven Intervention Framework for Improving Adherence to Growth Hormone Therapy Based on Clustering Analysis and Traffic Light Alerting Systems

Matheus ARAÚJO[a,1], Paula VAN DOMMELEN[b], Jaideep SRIVASTAVA[a] and Ekaterina KOLEDOVA[c]
[a]*Computer Science Department, University of Minnesota, Minneapolis, MN, USA*
[b]*The Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands*
[c]*Global Medical Affairs Cardiometabolic & Endocrinology, Merck Healthcare KGaA, Darmstadt, Germany*

Abstract. Recombinant human growth hormone (r-hGH) is an established therapy for growth hormone deficiency (GHD); yet, some patients fail to achieve their full height potential, with poor adherence and persistence with the prescribed regimen often a contributing factor. A data-driven clinical decision support system based on "traffic light" visualizations for adherence risk management of patients receiving r-hGH treatment was developed. This research was feasible thanks to data-sharing agreements that allowed the creation of these models using real-world data of r-hGH adherence from easypod™ connect; data was retrieved for 11,015 children receiving r-hGH therapy for ≥180 days. Patients' adherence to therapy was represented using four values (mean and standard deviation [SD] of daily adherence and hours to next injection). Cluster analysis was used to categorize adherence patterns using a Gaussian mixture model. Following a traffic lights-inspired visualization approach, the algorithm was set to generate three clusters: green, yellow, or red status, corresponding to high, medium, and low adherence, respectively. The area under the receiver operating characteristic curve (AUC-ROC) was used to find optimum thresholds for independent traffic lights according to each metric. The most appropriate traffic light used the SD of the hours to the next injection, with an AUC-ROC value of 0.85 when compared to the complex clustering algorithm. For the daily adherence-based traffic lights, optimum thresholds were >0.82 (SD, <0.37), 0.53–0.82 (SD, 0.37–0.61), and <0.53 (SD, >0.61) for high, medium, and low adherence, respectively. For hours to next injection, the corresponding optimum thresholds were <27.18 (SD, <10.06), 27.18–34.01 (SD, 10.06–29.63), and >34.01 (SD, >29.63). Our research indicates that implementation of a practical data-driven alert system based on recognised traffic-light coding would enable healthcare practitioners to monitor sub-optimally-adherent patients to r-hGH treatment for early intervention to improve treatment outcomes.

Keywords. Adherence, recombinant human growth hormone, growth hormone deficiency, cluster modeling, pediatrics

[1] Corresponding author, Matheus Araújo, Computer Science Department, University of Minnesota, Minneapolis, MN 55455, USA; E-mail: arauj021@umn.edu.fauc

1. Introduction

The use of r-hGH therapy to improve growth outcomes in children is well-established [1]. However, long-term studies have reported that patients often fail to achieve their full height potential [2], with sub-optimal adherence to r-hGH medication and poor persistence with the prescribed regimen considered major contributing factors [1, 3].

Historically, FAIR Principles [4] have relied upon clinical databases. However; there is potential to create new data-driven applications from patient-generated data [5]. In this study, we show how using the principle of responsible data-sharing [6] allowed creation of new innovations in data-driven applications for connected medical devices.

Digital health devices that monitor treatment adherence have the potential to improve patient/caregiver engagement and clinical outcomes [7]. For patients receiving r-hGH treatment (somatropin; Saizen®, Merck Healthcare KGaA, Darmstadt, Germany), the easypod™ auto-injector device, in combination with easypod™ connect, allows automatic recording and real-time data transmission of the date, time, and dose injected [8] enabling healthcare professionals to monitor patient adherence and growth outcomes. Data from connected devices has also contributed to the development of machine-learning algorithms to predict adherence behavior in multiple therapy areas [9-11].

Traffic light visualizations can help to guide and improve clinical decisions through the use of an established association between colors and related therapy signals which have been applied in multiple therapy areas [12, 13]. The concept of patient management through traffic light coding has also been successful in an emergency room setting, where the use of a three-tier urgency code helped prioritize patients for intervention [14].

2. Methods

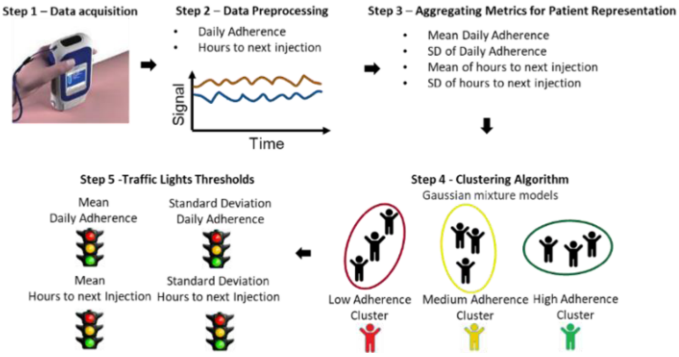

Step 1 – Data acquisition
Step 2 – Data Preprocessing
• Daily Adherence
• Hours to next injection
Signal
Time
Step 3 – Aggregating Metrics for Patient Representation
• Mean Daily Adherence
• SD of Daily Adherence
• Mean of hours to next injection
• SD of hours to next injection
Step 4 - Clustering Algorithm
Gaussian mixture models
Low Adherence Cluster
Medium Adherence Cluster
High Adherence Cluster
Step 5 -Traffic Lights Thresholds
Mean Daily Adherence
Standard Deviation Daily Adherence
Mean Hours to next Injection
Standard Deviation Hours to next Injection

Figure 1. Methodological overview (with each step referred to hereafter).

Adherence data (date and time of injection, injected dose, prescribed dose) from easypod™ connect were collected from 11,015 children receiving r-hGH for ≥180 days from January 2007–June 2019 (Figure 1, Step 1). Individual daily adherence was calculated: daily injected dose/daily prescribed dose (Figure 1, Step 2).

To determine injection consistency based on timing, 'hours to next injection' was computed using two daily signals (adherence and hours to next injection), aggregated with two metrics, the mean and SD of each patient during 180 days of r-hGH treatment. Patients' adherence to therapy was represented using a total of four values (mean and SD of adherence, and mean and SD of hours to next injection) (Figure 1, Step 3).

2.1. Model design and definition of adherence thresholds

Cluster analysis was used to categorize adherence patterns using a Gaussian mixture model, implemented in the Python 3.7 library, scikit-learn 0.23.1 [15]. Patient data were normalized using a z-score normalization (Standard Scaler) for each metric. To define a traffic light-based system, three clusters were assigned in the clustering algorithm under the assumption that, if three patient groups were distinguishable, adherence patterns between clusters would be well-defined. Thus, each cluster should present a considerably different mean adherence level followed by different values of adherence SD and mean, and SD of hours to next injection (Figure 1, Step 4).

In total, four traffic lights were computed, one for each aggregated feature per patient, for which, two thresholds were used to define green, yellow, or red alert status corresponding to high, medium, and low adherence, respectively. To define the pair of thresholds (mean and SD) for each feature, we generated 10,000 synthetic samples from the fitted Gaussian distribution, representing patients that cover the model's feature space: 6,248, high adherence; 2,824, medium adherence; and 928, low adherence. For each feature/traffic light, 50 possible thresholds from values of the synthetic samples were randomly selected. All possible pair-wise combinations of thresholds were evaluated to define the three clusters previously defined by the more complex clustering algorithm. The best pair of thresholds represent when to signal green, yellow, and red for each feature. The area under the receiver operating characteristic curve (AUC-ROC) was used to determine optimum thresholds based on higher values (Figure 1, Step 5).

3. Results

The mean age of patients was 10.2 years (SD, 3.1); 58% were male and 42% female.

3.1. Defined adherence clusters

A Kruskal-Wallis H-Test was performed which ensured the metrics significantly differed across clusters (p<0.01). Cluster distribution is presented in Table 1.

Table 1. Cluster centroids based on daily adherence and hours to next injection

Recorded Signal	Aggregating Metric	Cluster Centroids		
		High (n=6810, 62%)	**Medium (n=3186, 29%)**	**Low (n=1019, 9%)**
Daily adherence	Mean	0.92 (0.07)	0.72 (0.13)	0.46 (0.32)
	Standard deviation	0.32 (0.17)	0.48 (0.10)	0.47 (0.21)
Hours to next injection	Mean	25.58 (1.45)	28.38 (2.21)	42.84 (22.72)
	Standard deviation	5.41 (2.80)	14.93 (6.86)	196.58 (309.48)

3.2. Traffic light thresholds

The traffic light threshold values that best categorized patients according to high, medium, and low adherence for each metric are shown in Table 2. Patients with mean daily adherence of >0.82 trigger the green light, 0.53–0.82 the yellow light, and <0.53 the red light. In another example, considering the mean of hours to next injection, the best thresholds were 27.18 and 34.01, where values of <27.18 trigger the green light, 27.18–34.01 the yellow light, and >34.01 the red light. Table 2 also presents the AUC-

ROC score to verify the performance of each traffic light while defining the three groups. The highest AUC-ROC value was 0.85 for the SD of the hours to next injection; thus, this is the traffic light that most resembles the prediction of the clustering algorithm.

Table 2. Traffic light thresholds and their corresponding discriminating capability (AUC-ROC)

Recorded Signal	Aggregating Metric	Traffic Lights Thresholds			AUC-ROC
		Low (Red)	Medium (Yellow)	High (Green)	
Daily adherence	Mean	<0.53	0.53–0.82	>0.82	0.82
	Standard deviation	>0.61	0.37–0.61	<0.37	0.68
Hours to next injection	Mean	>34.01	27.18–34.01	<27.18	0.80
	Standard deviation	>29.63	10.06–29.63	<10.06	0.85

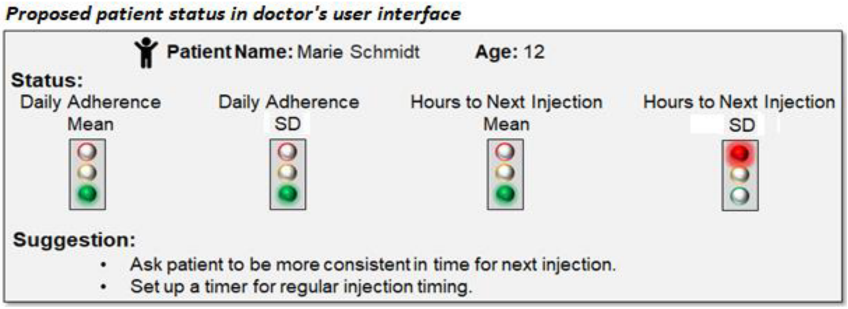
4. Discussion

This work is aligned with the need for further research into the visualization of data-driven applications [16]. Considering the cluster centroids detected by the model, higher SDs of both daily adherence and hours to next injection metrics were found in the lower adherence groups, suggesting that consistency is a key factor in distinguishing high and low adherence groups. Moreover, the high adherence group had an average mean daily usage of 0.92 (SD 0.07), reinforcing previous findings which classified 0.85% as the threshold defining good adherence to r-hGH therapy [3].

From the four proposed traffic lights, the SD of hours to next injection and mean daily adherence had a higher AUC-ROC. Thus, administering injections around the same time each day plays an essential role in maintaining high adherence to treatment and could serve as an important traffic light to alert clinicians to have discussions with patients/caregivers to mitigate the risk of sub-optimal adherence and, consequently, improve therapy effectiveness [2].

Study limitations include the small set of features used to create the traffic lights. Additional aggregating metrics, longitudinal data, and demographic information could be used but would also result in additional traffic lights. A filtering mechanism may be necessary, highlighting only the most useful traffic lights. It is recognized that other clustering algorithms (e.g. deep learning and time-series-based clustering algorithms) could be applied but with reduced model interpretability. Future work should evaluate whether the traffic lights would be meaningful for time frames beyond ≥180 days.

Figure 2 shows the proposed application of the traffic lights, where three out of four traffic lights correspond to good therapy use. However, the patient data triggers a red light for 'Hours to next injection SD'; and investigation is strongly recommended.



Proposed patient status in doctor's user interface

Patient Name: Marie Schmidt **Age:** 12

Status:

Daily Adherence Mean | Daily Adherence SD | Hours to Next Injection Mean | Hours to Next Injection SD

Suggestion:

- Ask patient to be more consistent in time for next injection.
- Set up a timer for regular injection timing.

Figure 2. Fictional example showing applicability of traffic light-based thresholds.

5. Conclusions

Our novel framework utilizes a clustering approach to categorize patients based on their r-hGH therapy adherence levels, as determined from data from easypod™ connect. The proposed traffic light alerting system serves as a practical tool for therapy personalization to support optimal growth outcomes. Additional formative human factor studies and subsequent validation studies with healthcare professionals are warranted to understand how far the presented traffic light system might support clinical decision making.

Disclosures

This research was made possible thanks to a collaboration between The Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands & Merck Healthcare KGaA, Darmstadt, Germany. MA, JS, and PvD have consultancy agreements with Merck. EK is an employee of Merck Healthcare KGaA, Darmstadt, Germany and holds shares in the company. Editorial assistance was provided by Sinéad Mutton, inScience Communications, Springer Healthcare Ltd, UK, funded by Merck Healthcare KGaA, Darmstadt, Germany.

References

[1] Graham S, et al. Identifying potentially modifiable factors associated with treatment non-adherence in paediatric growth hormone deficiency: A systematic review. Horm Res Paediatr. 2018;90(4):221-7.

[2] van Dommelen P, Koledova E, Wit JM. Effect of adherence to growth hormone treatment on 0-2 year catch-up growth in children with growth hormone deficiency. PLoS One. 2018;13(10):e0206009.

[3] Cutfield WS, Derraik JG, Gunn AJ, Reid K, Delany T, Robinson E, et al. Non-compliance with growth hormone treatment in children is common and impairs linear growth. PLoS One. 2011 Jan 31;6(1):e16223.

[4] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.

[5] Sayeed R, Gottlieb D, Mandl KD. SMART Markers: Collecting patient-generated health data as a standardized property of health information technology. NPJ Digit Med. 2020;3:9.

[6] Summary of Merck's responsible data sharing policy: https://www.merckgroup.com/research/healthcare/Responsible-Data-Sharing-Policy-EN.pdf.

[7] Bittner B, Schmit Chiesi C, et al. Connected drug delivery devices to complement drug treatments: potential to facilitate disease management in home setting. Med Devices (Auckl). 2019;12:101-27.

[8] Koledova E, Stoyanov G, Ovbude L, Davies PSW. Adherence and long-term growth outcomes: results from the easypod™ connect observational study (ECOS) in paediatric patients with growth disorders. Endocr Connect. 2018 Aug;7(8):914-23.

[9] Araujo A, Kazaglis L, Iber C, Srivastava J. A data-driven approach for continuous adherence predictions in sleep apnea therapy management, *2019 IEEE International Conference on Big Data,* pp. 2716-25.

[10] Araujo M, van Dommelen P, et al. Using Deep Learning for Individual-Level Predictions of Adherence with Growth Hormone Therapy. Stud Health Technol Inform. 2021 May 27;281:133-7.

[11] Tibble H, Chan A, Mitchell EA, Horne E, Doudesis D, Horne R, et al. A data-driven typology of asthma medication adherence using cluster analysis. Sci Rep. 2020 Sep 14;10(1):14999.

[12] Dagliati A, Sacchi L, Tibollo V, Cogni G, Teliti M, Martinez-Millana A, et al. A dashboard-based system for supporting diabetes care. J Am Med Inform Assoc. 2018 May 1;25(5):538-47.

[13] Saposnik G, Grueschow M, Oh J, Terzaghi MA, Kostyrko P, Vaidyanathan S, et al. Effect of an educational intervention on therapeutic inertia in neurologists with expertise in multiple sclerosis: A randomized clinical trial. JAMA Netw Open. 2020 Dec 16;3(12).

[14] Leppaniemi A, Jousela I. A traffic-light coding system to organize emergency surgery across surgical disciplines. Br J Surg. 2014 Jan;101(1):e134-40.

[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

[16] West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. J Am Med Inform Assoc. 2015 Mar;22(2):330-9.

doi:10.3233/SHTI210804

Development of a FHIR Layer on Top of the OMOP Common Data Model for the CAPABLE Project

Matteo GABETTA [a,1], Anna ALLONI [a], Francesca POLCE [b], Giordano LANZOLA [b], Enea PARIMBELLI [b] and Nicola BARBARINI [a]
[a] *BIOMERIS (BIOMEdical Research Informatics Solutions), Pavia, Italy*
[b] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy*

Keywords. OMOP, HL7 FHIR, API, Cancer

1. Introduction

The CAPABLE (CAncer PAtient Better Life Experience) project [1], funded in the H2020 program[2], is developing a novel system to improve the quality of life of cancer patients managed at home. CAPABLE system is based on a distributed software architecture where different components cooperate with the aim of "early detecting and managing cancer-related issues and at satisfying the needs of patients and their home caregivers". One of the core CAPABLE components is the "CAPABLE Data Platform" (DP); the main objective of the DP is to provide a persistent layer where to store and fetch all project's patient-related data.

To guarantee a state-of-the art level component, OMOP [2] Common Data Model (CDM) and HL7-FHIR [3] have been chosen for persistency and exchange format respectively. The main reason for choosing OMOP is due to its "Standardized clinical data" tables, which are designed to hold disparate patient-related data, and to the "Standardized vocabularies", a set of international standard terminologies which are consolidated into the same code system. Alongside with the need of having a standard model to represent persisted data, the project also needed a reliable format for exchanging those data in a web-service safe mode: FHIR was chosen for this purpose because it is based on a "composition approach", representing standard clinical entities as resources that can be combined with each other.

The *omoponfhir* open-source project [4] constituted the starting point for the development of the DP (which in fact can be considered a fork of this project); in this article we highlight the main changes/additions and customization of *omoponfhir* to make it fit the aims and requirements of the CAPABLE project.

[1] Corresponding Author, BIOMERIS, Via Ferrata 1, 27100 Pavia, Italy; E-mail: matteo.gabetta@biomeris.it.

[2] The work described in this article has been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 875052 - CAPABLE

2. Methods

The development of DP is covering several aspects of the *omoponfhir* project; the most relevant general improvements concern: (i) an easier management of FHIR *Coding* resource, that is used by mostly all other resources; (ii) the option of automatically translating any FHIR *Coding* to its standard OMOP synonym (if present), thus allowing the interoperability between different terminologies directly in FHIR; (iii) the possibility to go beyond a the flat searching model, where all constraints are evaluated in sequence, with the introduction of a hierarchical query model.

For what concerns the supported FHIR Resources, several new search parameters have been implemented in the code base; moreover, new resources, that are not managed by the original project, have been implemented: Goal, List and Communication.

Another relevant improvement is related to the FHIR *Observation* Resource, which can represent concepts that map multiple OMOP domains (e.g., Observation, Measurement, Procedure and Condition Occurrence): a fine-grained management has been implemented, taking also into account the fact that an Observation could be possibly negated or further specified by scales or grades.

The OMOP CDM has been also extended with new tables; the most relevant one is *f_update* and allows to specify the *lastupdated* attribute for all the facts that can be stored in the DP: this is a crucial information for the CAPABLE system and for FHIR in general that wasn't manageable with the standard OMOP CDM. Finally, the CTCAE (Common Terminology Criteria for Adverse Events) [5] is in the process of being added to the OMOP Vocabularies and mapped to standard OMOP concepts.

3. Preliminary Results and Next Steps

The overall CAPABLE system, DP included, has already undergone two development iterations each one concluded with a live demonstration. The demonstrations, designed together with clinicians and patients involved in the project, revolve around prototypical fictional patients, following them in different settings (enrollment visit, follow-up, homecare) and focus on specific clinical guidelines.

During the demonstrations, mostly all of the CAPABLE components have interacted with DP by storing and fetching more than 130 different FHIR Resources, which have been saved and then retrieved from the customized OMOP CDM.

Next steps involve the release in production of the DP to comply with the all the project's use cases.

References

[1] CAPABLE H2020 Project [Internet]. Capable. 2021 [cited 2021Jul26]. Available from: https://capable-project.eu/

[2] The OHDSI community. (2020). The Book of OHDSI (Version 2020-10-19). The Book of OHDSI (p. 470). Zenodo. http://doi.org/10.5281/zenodo.4265256

[3] FHIR v4.0.1 [Internet]. Hl7.org. 2021 [cited 26 July 2021]. Available from: https://www.hl7.org/fhir/

[4] Choi, M., Starr, R., Braunstein, M., & Duke, J. (2016). OHDSI on FHIR platform development with OMOP CDM mapping to FHIR Resources. In OHDSI Symposium, Observational Health Data Sciences and Informatics, Washington, DC.

[5] CTCAE [Internet]. NCI. 2021 [cited 26 July 2021]. Available from https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210805

Automatic Data Transfer from OMOP-CDM to REDCap: A Semantically-Enriched Framework

Emanuele GIRANI[a], Matteo GABETTA[b,1], Anna ALLONI[b], Morena STUPPIA[b], Lucia SACCHI[a] and Nicola BARBARINI[b]
[a] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy*
[b] *BIOMERIS (BIOMEdical Research Informatics Solutions), Pavia, Italy*

Keywords. OMOP, OHDSI, REDCap, eCRF, EDC, Real World Data

1. Introduction and Background

Clinical research is currently limited by the need to manually enter data related to clinical trials through EDC (Electronic Data Capture) or eCRF (electronic Case Report Form). This implies replication of data between Electronic Health Record systems (EHR) and eCRF, the employment of a considerable amount of time and resources and easily leads to errors. Thus, the systematic reuse of Real World Data (mainly EHR data) to automatically fill in eCRF may represent a turning point in clinical trials [1].

OMOP/OHDSI (Observational Medical Outcomes Partnership - Observational Health Data Sciences and Informatics) is an ideal middleware to be interposed between EHR and eCRF, in order to decouple the complexity of the clinical sources from the target eCRF [2]. Its use of both a standardized data model and the main standardized terminologies, makes it a particularly suitable candidate.

REDCap (Research Electronic Data CAPture) is a widely adopted web-based eCRF system for non-profit studies [3]. There are different solutions to automatically import data into a REDCap study (direct ETL through REDCap API, Dynamic Data Pull (DDP), Clinical Data Pull (CDP)), but to the best of our knowledge no specific project is focusing on automatic data transfer from OMOP Common Data Model (CDM) to REDCap.

The aim of this work is proposing a semantically-enriched framework to support automatic data transfer from OMOP CDM to REDCap eCRF.

2. Method

The main components of the system are: (i) a semantic annotation framework to extend a REDCap study's metadata and, (ii) a software component which actually operates the data transfer accordingly to the semantic annotations. The semantic annotation is added

[1] Corresponding Author, BIOMERIS, Via Ferrata 1, 27100 Pavia, Italy; E-mail: matteo.gabetta@biomeris.it.

to the REDCap Data Dictionary, a specific metadata that annotates the study's elements (i.e., variables); these annotations are written in JSON format with different sections. The first section identifies, among the different terminologies included in the OMOP vocabularies, which concept better describes the element itself (e.g., SNOMED code for "Gender") and in some cases the possible values (e.g., SNOMED concepts for "Female"). Other sections of the semantic annotation cover aspects related to defining how the data transfer process shall occur considering several aspects: (i) the reference date, in the REDCap study, associated to a specific time-dependent element (e.g., a laboratory value), to accurately fetch the OMOP database; (ii) a time tolerance to relax the OMOP queries (e.g. reference date ± 1 day); (iii) how dependent elements (e.g. because of a branching logic rule) have to be treated by the system; (iv) the actual order in which the data transfer attempts for a single or groups of variables have to be executed.

Besides the semantic annotation of the study, we developed a tool in Java language which: (i) reads the semantic annotation through the REDCap standard API, (ii) has access to an external lookup table to match REDCap record IDs with OMOP IDs, (iii) allows to execute the data transfer according to different policies/schedules, for example, on a periodic base (e.g., once a day, every night), whenever a new patient is created in REDCap (leveraging the trigger functionality of REDCap) or on-demand (i.e., the user of the tool decides when to fetch data).

3. Preliminary Results and Discussion

A first prototype of the framework has been developed in Java to test the proposed approach. We decided to test the framework on the Italian Registry for Severe Asthma Patients (SANI). SANI is an ideal test case for the system because it is managed in REDCap and an updated OMOP version of its data is maintained by the ERS SHARP initiative, where a particular subset of 198 OMOP concepts has been chosen as the CDM for several European registries for severe asthma in order to build a federated network.

In particular we are testing the framework on 148 REDCap variables representing different input format (numbers, dropdown, etc.) and data types (demographics, anamnesis, tests and therapies). Some preliminary results show a promising ability of the framework to automatic fill in SANI REDCap eCRF from the current SHARP-SANI OMOP database. These results foreshadow a great saving in terms of time and resources for data collection activity in REDCap eCRF starting from a OMOP CDM.

References

[1] Griffon N, Pereira H, Djadi-Prat J, García MT, Testoni S, Cariou M, Hilbey J, N'Dja A, Navarro G, Gentili N, Nanni O, Raineri M, Chatellier G, Gómez De La Camara A, Lewi M, Sundgren M, Daniel C, Garvey A, Todorovic M, Ammour N. Performances of a Solution to Semi-Automatically Fill eCRF with Data from the Electronic Health Record: Protocol for a Prospective Individual Participant Data Meta-Analysis. Stud Health Technol Inform. 2020 Jun 16;270:367-371.

[2] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.

[3] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)-A Metadata-driven Methodology and Workflow Process for Providing Translational Research Informatics Support. Journal of Biomedical Informatics. 2009 Apr; 42(2):377–381.

This page intentionally left blank

Section II

Methods for the Adoption of FAIR Principles & Privacy and Security Aspects Applying FAIR in Health Research

This page intentionally left blank

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210807

FAIRification Efforts of Clinical Researchers: The Current State of Affairs

Martijn G. KERSLOOT [a,b,1], Philip VAN DAMME [a], Ameen ABU-HANNA [a], Derk L. ARTS [b] and Ronald CORNET [a]

[a] *Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands*
[b] *Castor EDC, Amsterdam, The Netherlands*

Abstract. The FAIR Principles are supported by various initiatives in the biomedical community. However, little is known about the knowledge and efforts of individual clinical researchers regarding data FAIRification. We distributed an online questionnaire to researchers from six Dutch University Medical Centers, as well as researchers using an Electronic Data Capture platform, to gain insight into their understanding of and experience with data FAIRification. 164 researchers completed the questionnaire. 64.0% of them had heard of the FAIR Principles. 62.8% of the researchers spent some or a lot of effort to achieve any aspect of FAIR and 11.0% addressed all aspects. Most researchers were unaware of the Principles' emphasis on both human- and machine-readability, as their FAIRification efforts were primarily focused on achieving human-readability (93.9%), rather than machine-readability (31.2%). In order to make machine-readable, FAIR data a reality, researchers require proper training, support, and tools to help them understand the importance of data FAIRification and guide them through the FAIRification process.

Keywords. FAIR data, medical research, Research Data Management

1. Introduction

In order to improve and support the reuse of scholarly output, a multidisciplinary group of researchers published the fifteen FAIR Guiding Principles for scientific data management and stewardship in 2016, stating that scholarly output should be Findable, Accessible, Interoperable, and Reusable, both for machines and for people [1]. The paper describing the Principles explicitly emphasizes FAIRness for machines, and in another publication, Mons et al. reiterate that "FAIR is not just about humans being able to find, access, reformat and finally reuse data". Researchers are increasingly required, either by their institutions or funders, to FAIRify their data (i.e., to make their data more FAIR). However, FAIR provides guidance for research data management and is not a standard [2], hence researchers need to interpret the Principles for their use case and make implementation choices accordingly [3].

Within the biomedical community, there are various initiatives that endorse the Principles and aim to develop guidance and tools for researchers [4]. Sinachi et al., for ex-

[1] Corresponding Author: Martijn G. Kersloot, m.g.kersloot@amsterdamumc.nl.

ample, developed a FAIRification workflow specific to health research [5]. Practical examples of the implementation of the FAIR Principles include the collection of linked, human- and machine-readable data of COVID-19 patients [6] and rare disease patients [7,8] and the reuse of such machine-readable data to perform distributed analyses [9]. Despite these initiatives, little is known about the knowledge and efforts of individual researchers regarding data FAIRification. Therefore, we sought to gain insight into clinical researchers' understanding of and experience with data FAIRification.

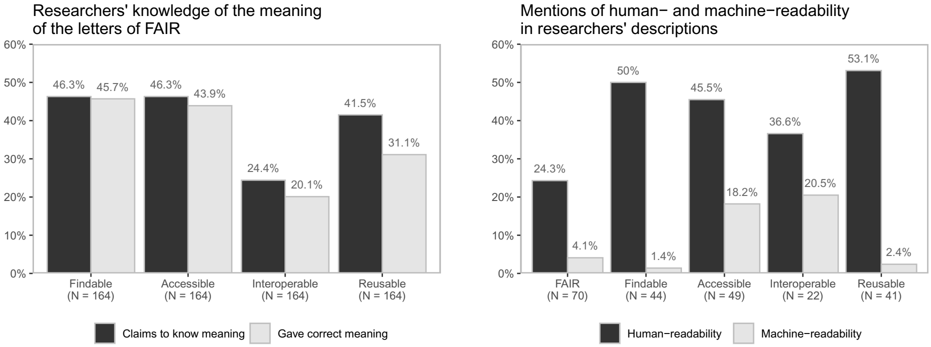
2. Methods

An online English questionnaire was developed [10]. In the questionnaire, researchers were asked to report if they were familiar with the FAIR Principles, if they knew the meaning of each of the letters of FAIR (Findability, Accessibility, Interoperability, and Reusability), and if they could describe what each aspect of FAIR entailed. In addition, they were asked to rank (no effort, very little effort, some effort, a lot of effort) and describe their current efforts for making data more FAIR. Clinical researchers were invited by email by Research Data Management departments in five out of the seven Dutch University Medical Centers (UMCs). PhD student associations of three UMCs sent out invitation emails to their members. In addition, users of Electronic Data Capture (EDC) platform Castor EDC [11] received an invitation via an popup in the platform. The invitation did not include any mention of FAIR, to ensure that potential participants that did not know about FAIR were also included. Researchers were eligible to participate in our study if they were setting up databases for clinical research. Consent of respondents was required before the questionnaire could be opened. Data were collected in Castor EDC [11] between November 27, 2020, and February 27, 2021. Statistical analyses were performed using R (version 4.0.2) [12]. Questionnaires with complete answers to mandatory questions were included in the analysis. Free-text answers given by researchers describing their interpretation of the FAIR Principles and their FAIRification efforts were assessed by two medical informaticians (MK, PvD). We assessed whether the descriptions of the Principles were related to human or machine readability. The FAIRification efforts were divided into categories by way of induction, and the contribution of each category to human and machine readability was evaluated.

3. Results

A total of 164 researchers completed the questionnaire. Demographics can be found at [10]. 51.2% (n = 84) opened the questionnaire via an invitation in a UMC and 48.8% (n = 80) did so via the popup in the EDC system. The majority (87.2%, n = 143) worked in a UMC and was a PhD candidate (84.8%, n = 139). 64.0% (n = 105) of all researchers had heard of the FAIR Principles, 45.1% (n = 74) claimed to know what the Principles entailed, and 56.1% (n = 72) claimed to know the meaning of at least one of the letters in FAIR (Findable, Accessible, Interoperable, and Reusable). Not all researchers that claimed to know the meaning of one or more of the letters, gave a correct meaning (Figure 1.1). Findable was the aspect that was understood best, followed by Accessible, Reusable, and Interoperable. The minority of the descriptions of the Principles and aspects given by researchers were related to machine-readability (Figure 1.2).

After explaining the Principles to the researchers, 81.1% (n = 133) of them stated that they have spent at least very little effort to make their data more Findable, Accessible, Interoperable, or Reusable (*any* aspect of FAIR). For *all* aspects of FAIR, this was 25.6% (n = 42). 62.8% (n = 103) of the researchers spent some or a lot of effort to achieve *any* aspect of FAIR. For *all* aspects this was 11.0% (n = 18).



(1) (2)

Figure 1. Researchers' knowledge and descriptions of the individual FAIR aspects

88.0% (n = 117) of the researchers that spent at least very little effort in making their data FAIR provided a description of their FAIRification efforts. An overview of the most common reported efforts is listed in Table 1. Of all mentioned efforts (N = 231), 93.9% (n = 217) focused on human-readability (e.g., using a data dictionary or using social media to share research outputs) and 31.2% (n = 72) focused on machine-readability (e.g., submitting data and metadata to a data repository or using standardized terminologies).

Table 1. Efforts to make data more FAIR, as reported by researchers

	F		A		I		R		Total	
	N = 72		N = 84		N = 64		N = 78		N = 117	
	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
Standardize data collection	5	6.9	1	1.2	23	35.9	12	15.4	41	35.0
Add dataset descriptions	5	6.9	6	7.1	4	6.3	17	21.8	32	27.4
Using a shared drive	2	2.8	19	22.6	0	0.0	1	1.3	22	18.8
Using a data capture system / EMR	6	8.3	9	10.7	4	6.3	2	2.6	21	17.9
Deposit data in a data archive	6	8.3	12	14.3	1	1.6	0	0.0	19	16.2
Create study protocol and register study	7	9.7	1	1.2	4	6.3	4	5.1	16	13.7
Provide data availability/access statement	8	11.1	6	7.1	0	0.0	1	1.3	15	12.8
Standardized file management	11	15.3	2	2.4	0	0.0	0	0.0	13	11.1
Add data as supplementary material	5	6.9	3	3.6	0	0.0	1	1.3	9	7.7
Storing data in a database	2	2.8	4	4.8	2	3.1	1	1.3	9	7.7
Use social media	5	6.9	2	2.4	0	0.0	1	1.3	8	6.8
Add documentation	1	1.4	0	0.0	0	0.0	6	7.7	7	6.0
Using a Data Management Plan	3	4.2	1	1.2	1	1.6	1	1.3	6	5.1

This table does not include efforts reported by less than 5% of the researchers.

4. Discussion

In this study, we conducted an online questionnaire to gain insights into clinical researchers' understanding of the FAIR Principles and their FAIRification experiences. We found that 64.0% (n = 105) of the researchers have heard of the Principles and that 62.8% (n = 103) of all researchers spent at least some effort to achieve to achieve *any* aspect of FAIR, and 11.0% (n = 18) regarding *all* aspects.

A strength of our study is that we invited researchers via an EDC platform and via UMCs, ensuring that we included researchers who were responsible for data management, and thus FAIRification. Moreover, we made certain that our invitation did not mention FAIR, in order to recruit both researchers who were and were not familiar with the Principles. Unfortunately, we were unable to verify the background of the respondents, because potential respondents were not directly invited by the research team to comply with the General Data Protection Regulation. Lastly, partially filled-in questionnaires were excluded from the analysis, which could have introduced selection bias, for researchers who were unaware of FAIR could have stopped filling in the questionnaire.

The definition of the Principles and meaning of the individual letters of FAIR were known only by a minority of the researchers participating in our study and only a small amount of researchers spent effort to achieve FAIRness. Moreover, it is still unclear to researchers which efforts contribute to which aspect of FAIR. For example, using a data capture system to collect data does not specifically mean that the data collected in that system are more Findable (F), Accessible (A), or Interoperable (I) with data collected in other systems. The same applies to the use of standardized variable names for collecting data, where free-text variable names do not specifically make data more Interoperable (I) for machines. This indicates the need for simple, easy-to-follow explanations of what the Principles mean in practice and how they affect the (daily) work of clinical researchers. Specifically, researchers should receive training and guidance to help them understand which steps they should take to make their data more FAIR and how these steps are related to the individual aspects of the FAIR Principles. A first step toward achieving this is to ensure that there is convergence on FAIR implementations in the healthcare domain. Community-specific FAIR Implementation Profiles [13] can drive this convergence.

Most descriptions of researchers focused on human-readability, rather than machine-readability, and only a minority of the researchers focused on achieving machine-readability. This might be due to the fact that the current workflows mentioned in the literature and software available to make (meta)data machine-readable require a significant amount of background knowledge (e.g., knowledge of data modeling, terminology systems, or metadata schemes). Researchers should move away from the use of such "professorware" and turn to sustainable systems [14]: professional products and services that support them in the creation and use of FAIR data [15]. To ensure that the systems are integrated into the researchers' existing working processes, research software vendors should develop these systems in collaboration with researchers and research support staff, and integrate them with systems that are currently used.

5. Conclusion

A large number of clinical researchers is currently unaware of the definition of the FAIR Principles and their emphasis on both human- and machine-readability. Researchers are undertaking efforts to make their data more FAIR, but their focus is primarily on human-readability, rather than machine-readability. In order to make machine-readable, FAIR data a reality, researchers need proper training, support, and tools to help them understand the importance of data FAIRification and guide them through the FAIRification process.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016 Mar;3(1). Available from: https://doi.org/10.1038/sdata.2016.18.

[2] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017 Mar;37(1):49–56. Available from: https://doi.org/10.3233/ISU-170824.

[3] Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al.. FAIR Principles: Interpretations and Implementation Considerations. MIT Press - Journals; 2020. Available from: https://doi.org/10.1162/dint_r_00024.

[4] Trifan A, Oliveira JL. Towards a More Reproducible Biomedical Research Environment: Endorsement and Adoption of the FAIR Principles. In: Biomedical Engineering Systems and Technologies. Springer International Publishing; 2020. p. 453–470. Available from: https://doi.org/10.1007/978-3-030-46970-2_22.

[5] Sinaci AA, Núñez-Benjumea FJ, Gencturk M, Jauer ML, Deserno T, Chronaki C, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of Information in Medicine. 2020 Jun;59(S 01):e21–e32. Available from: https://doi.org/10.1055/s-0040-1713684.

[6] Reisen M, Oladipo F, Stokmans M, Mpezamihgo M, Folorunso S, Schultes E, et al. Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced Genetics. 2021 Jun;2(2). Available from: https://doi.org/10.1002/ggn2.10050.

[7] Jannik S, Dennis K, Jens G, Christian-Alexander B, Marco R, van Enckevort David, et al. OSSE Goes FAIR - Implementation of the FAIR Data Principles for an Open-Source Registry for Rare Diseases. Studies in Health Technology and Informatics. 2018;253:209–213. Available from: https://doi.org/10.3233/978-1-61499-896-9-209.

[8] Groenen KHJ, Jacobsen A, Kersloot MG, dos Santos Vieira B, van Enckevort E, Kaliyaperumal R, et al. The de novo FAIRification process of a registry for vascular anomalies. Orphanet Journal of Rare Diseases. 2021 Sep;16(1). Available from: https://doi.org/10.1186/s13023-021-02004-y.

[9] Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intelligence. 2020 Jan;2(1-2):96–107. Available from: https://doi.org/10.1162/dint_a_00032.

[10] Kersloot MG, van Damme P, Abu-Hanna A, Arts DL, Cornet R. FAIRification efforts of clinical researchers: the current state of affairs. figshare; 2021. Available from: https://doi.org/10.6084/m9.figshare.c.5617396.v1.

[11] Castor EDC. Castor Electronic Data Capture; 2020. Available from: https://www.castoredc.com.

[12] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

[13] Schultes E, Magagna B, Hettne KM, Pergl R, Suchánek M, Kuhn T. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Lecture Notes in Computer Science. Springer International Publishing; 2020. p. 138–147. Available from: https://doi.org/10.1007/978-3-030-65847-2_13.

[14] Mons B. Data Stewardship for Open Science. Chapman and Hall/CRC; 2018. Available from: https://doi.org/10.1201/9781315380711.

[15] van Vlijmen H, Mons A, Waalkens A, Franke W, Baak A, Ruiter G, et al. The Need of Industry to Go FAIR. Data Intelligence. 2020 Jan;2(1-2):276–284. Available from: https://doi.org/10.1162/dint_a_00050.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210808

Best Research Practice Implementation: The Experience of the N.N. Burdenko National Medical Research Center of Neurosurgery

Gleb DANILOV[a,1] , Michael SHIFRIN[a], Yulia STRUNINA[a], Timur ISHANKULOV[a], Timur ZAGIDULLIN[a], Elizaveta MAKASHOVA[a], Igor PRONIN[a], Nikolay KONOVALOV[a] and Alexander POTAPOV[a]

[a] *Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center of Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. Implementing the best research principles initiates an important shift in clinical research culture, improving efficiency and the level of evidence obtained. In this article, we share our own view on the best research practice and our experience introducing it into the scientific activities of the N.N. Burdenko National Medical Research Center of Neurosurgery (Moscow, Russian Federation). While being adherent to the principles described in the article, the percentage of publications in the international scientific journals in our Center has increased from 7% to 27%, with an overall gain in the number of articles by 2 times since 2014. We believe it is important that medical informatics professionals equally to medical experts involved in clinical research are familiar with the best research principles.

Keywords. Best research practice, neurosurgery, data management, FAIR, biostatistics

1. Introduction

Medicine is traditionally an area entirely focused on practice and application: providing the most effective, safe, and cost-effective care for people with health issues. Probably due to the complexity of human biology and our limited knowledge of it, medicine is difficult to formalize. Many concepts do not have clear and generally accepted definitions. Many definitions are superficial and rely only on the explicit pathology manifestation. At the same time, the concept of evidence-based medicine, which has gained a deserved popularity and recognition in recent decades, postulates scientific verification of the effectiveness, safety, and economic feasibility of medical interventions as the basis for decision-making in medicine [2]. Scientific hypothesis testing inevitably relies on formalization, logic, and rigorous operational definitions, essential for exact sciences. Due to the specificity of training and practice in medicine, these approaches are not at the forefront in the system of a doctor's professional thinking. The culture of well-designed research is not taken for granted in medicine. The research activity itself requires serious multidisciplinary support, in which the role of medical

[1] Corresponding author, Gleb Danilov, the N.N. Burdenko National Medical Center of Neurosurgery, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

informatics is of vital importance. The quality of research outcome cannot rest solely with doctors. IT specialists are much responsible for handling data, which necessitates a good understanding of clinical research ethics, regulation, and methodology.

The pharmaceutical industry has developed best practices in human research over the years. We have to admit that research opportunities tend to be more restrained in the academic field than in big pharma. However, in our opinion, implementing best research principles is a cultural shift that requires not so much money as changing the way of thinking. In this article, we want to share our own experience of introducing the best research practices into scientific activities of the N.N. Burdenko National Medical Research Center of Neurosurgery (the NSC, former the N.N. Burdenko Neurosurgery Institute, Moscow, Russian Federation).

2. Rationale for best research practice implementation in medical research

The NSC is a leading neurosurgical institution in the Russian Federation, one of the largest neurosurgical facilities in the world. The National Neurosurgery Center has 300 beds and annually performs up to 10,000 neurosurgical interventions with a postoperative mortality rate of less than 0.5%. For 21 years of medical information systems operation, the National Center for Neurosurgery has accumulated unique and large data archives, mining of which is absolutely justified but not easy. The data generation speed in high-tech neurosurgery exceeds the ability to extract knowledge from data. Thus, the strategic goals for the development of the NSC research activities are:

- to improve the level of research quality and evidence due to best research practices
- to increase the efficiency of the secondary data use in scientific research
- to improve data quality being collected in research
- to increase the likelihood of papers acceptance in high-ranked journals
- to increase the number of citations

To achieve these goals, it seems appropriate to influence the complex research process at its consecutive stages.

3. Key research principles we consider

Below we list the main ideas we concentrated on to improve our research practice. These are *research planning, biostatistics, data management, reproducible statistical analysis, medical writing and project management.* All the research principles we enlist follow the good clinical practice (GCP) statement, an international ethical and scientific quality standard for designing, recording, and reporting research with human subjects involvement. We consider sticking to this standard obligatory. That is why we believe the experts in medical informatics should invest a certain time to get acquainted with the ICH GCP document [4].

3.1. Research planning

Good research practice starts with resolving ethics and legal issues. Is it legal and moral to do a certain study in humans? Does it carry more benefits than risks for patients? Is it possible and necessary to protect or insure the participants from unfavorable events related to a study? One should answer all these questions before the study starts taking into account national and global regulations.

The technical part of research begins with a well-thought plan written normally in the form of a protocol. Writing a rigorous research plan sounds like a well-known principle, nevertheless commonly underestimated and even regularly ignored. Writing a research protocol could be intently addressed in a series of articles - this is such an important principle. It completely determines all subsequent stages of the research process and, therefore, the success in achieving the results, the likelihood of publication, and any other "return on investment." Planning a study provides an unambiguous goal statement, helps minimize common errors and biases in research aim, design choice, patient selection, data collection, data analysis, interpretation and presentation of results, etc.

At the NSC we developed a protocol template which includes the following sections:

- title page with main project identifiers,
- project team,
- definitions and abbreviations,
- project goals and tasks,
- population with inclusion/exclusion criteria,
- rationale/background for the study,
- research design and methodology with project scheme,
- randomization plan (when necessary),
- primary and secondary end points,
- data management plan,
- statistical analysis plan,
- case report form (CRF),
- additional resources,
- informed consent,
- investigator commitments,
- local advisory board approvement,
- local ethics board approvement,
- references.

We consider filling in these sections is necessary to ensure the best research practice in the upcoming research procedures as requires by GCP statement.

Certain mnemonic rules facilitate the start of writing a protocol (such as PICO – Patient, Intervention, Comparator, Outcome). We believe in doing the following first when planning a project:

- Summarizing known (via systematic literature search)
- Defining unknown
- Asking a proper research question
- Defining the right research hypothesis

The goal statement, design, data collection scheme come accordingly to the main hypothesis.

3.2. Biostatistics

This principle is overrated and underestimated simultaneously. Medical experts often believe that the fact of statistical analysis itself determines the study's success. Yet, statistical analysis is secondary to the correct task formulation and the quality of the collected data. Biostatistician participates in study planning and writing a research protocol, translates the main hypothesis into a formalized version which is testable with a statistical approach, chooses research design, defines the set of data to collect, picks up the statistical tests, calculates sample sizes to guarantee research power, performs data analysis. It is quite clear that without biostatistics, evidence-based medicine has no foundation. A biostatistician is a trained specialist a doctor cannot fully replace, even if the latter has mastered certain statistical analysis methods. We believe it is important to involve such a specialist in a clinical research team.

3.3. Data management

Data management is a key process aimed to ensure the unambiguity, completeness, security, and reliability of data collection and storage in research. The efficiency and quality of the data management are assured with adherence to GCP using clinical data management systems (CDMS). At the NSC, a stand-alone version of the CDMS REDCap is used as a data management tool in studies planned since 2017 [1,3]. This system provides data entry organization in electronic CRF, a unified structure, format, integrity, security, availability, verification, and quality control of data, which is important for the reproducibility of statistical analysis, information security, and multicenter distributed research, final research quality. In fact, the system contributes to the implementation of the FAIR principles into clinical research [5].

3.4. Reproducible statistical analysis

The statistical analysis becomes reproducible with code written in programming languages. Compared with push-button interface in statistical software, scripts are beneficial in reproducibility when typical data analysis must be repeated many times as data accumulates. Reporting the analysis procedure in code is another quality control tool. At the NSC, we typically use R for common statistical analysis and normally code in Python for machine learning projects.

3.5. Medical writing

Comments on the quality of scientific English text written by non-native speakers are typical from journal reviewers and editors. However, writing a text in proper English is only part of the task. Complete and clear reporting of research results is no less important for their subsequent correct interpretation. The EQUATOR (Enhancing the QUAlity and Transparency Of health Research) Network is an international initiative to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines [6]. Nowadays, many medical journals require following these guidelines when submitting a manuscript. However, we strongly advise adopting them in every research report as a best practice.

3.6. Project management

No task can be accomplished without control and sometimes pushing. Even the most well-planned research can come to naught without good management, especially when many stakeholders are involved. The "research orchestra" conductor is a project manager who understands when different activities must be completed and controls the overall movement towards the goal.

4. The results of best practice implementation

Shifting towards a new research culture with adherence to the above-mentioned principles has influenced the productivity of scientific research in our Center in recent years. Since 2017, data for almost 60 scientific projects have been managed in REDCap. Since 2018, research protocols are developed for almost all new projects. Although this research paradigm is instilled gradually and not all principles are well perceived by medical experts, the percentage of publications in the international peer-reviewed scientific journals has increased from 7% (8 of 111) to 27% (60 of 222), with an overall gain in the number of scientific articles by 2 times since 2014.

5. Conclusion

High-quality research, regardless of its objects and subjects, can be ensured by adhering to a set of principles and engaging a multidisciplinary team with dedicated competencies in research planning, data collection and analysis, and project management. We believe it is important that medical informatics professionals equally to medical experts involved in clinical research are familiar with these principles.

Our work was supported by Russian Foundation for Basic Research (grant 18-29-01052).

References

[1] Danilov GV, Shifrin MA, Strunina YV, Pronkina TE, Ishankulov TA, Burov AA, Dorofeyuk YA, and Potapov AA. Clinical Research Data Management: Experience of the N.N. Burdenko NMRC of Neurosurgery [in Russian]. Vrach i Inf. Tehnol. 2020; 6–14.

[2] Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. The Lancet. 2017 Jul 22;390(10092):415-23.

[3] Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN. The REDCap consortium: Building an international community of software platform partners. Journal of biomedical informatics. 2019 Jul 1;95:103208.

[4] INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE (ICH) ICH HARMONISED GUIDELINE INTEGRATED ADDENDUM TO ICH E6(R1): GUIDELINE FOR GOOD CLINICAL PRACTICE E6(R2), (2016). https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf (accessed July 30, 2021).

[5] FAIR Principles - GO FAIR, (n.d.). https://www.go-fair.org/fair-principles/ (accessed July 30, 2021).

[6] The EQUATOR Network | Enhancing the QUAlity and Transparency Of Health Research, (n.d.). https://www.equator-network.org/ (accessed July 30, 2021).

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210809

Pilot Study of an e-Cohort to Monitor Adverse Event for Patient with Hip Prostheses from Clinical Data Warehouse

Thibault DHALLUIN[a,b,1], Marie ANSOBORLO[a,b], Philippe ROSSET[b,c], Hervé THOMAZEAU[d], Marc CUGGIA[e] and Leslie GUILLON[a,b]
[a] *Department of Medical Information, University Hospital of Tours, Tours, France*
[b] *Medical school, University of Tours, EA 7505 EES, Tours, France*
[c] *Department of Orthopedic Surgery, University Hospital of Tours, Tours, France*
[d] *Federation of Orthopaedic Surgery, University Hospital Sud, 35203 Rennes, France*
[e] *Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France*

Abstract. Hip arthroplasty represents a large proportion of orthopaedic activity, constantly increasing. Automating monitoring from clinical data warehouses is an opportunity to dynamically monitor devices and patient outcomes allowing improve clinical practices. Our objective was to assess quantitative and qualitative concordance between claim data and device supply data in order to create an e-cohort of patients undergoing a hip replacement.

We performed a single-centre cohort pilot study, from one clinical data warehouse of a French University Hospital, from January 1, 2010 to December 31, 2019. We included all adult patients undergoing a hip arthroplasty, and with at least one hip medical device provided. Patients younger than 18 years or opposed to the reuse of their data were excluded from the analysis. Our primary outcome was the percentage of hospital stays with both hip arthroplasty and hip device provided. The patient and stay characteristics assessed in this study were: age, sex, length of stay, surgery procedure (replacement, repositioning, change, or reconstruction), medical motif for surgery (osteoarthritis, fracture, cancer, infection, or other) and device provided (head, stem, shell, or other).

We found 3,380 stays and 2,934 patients, 96.4% of them had both a hip surgery procedure and a hip device provided. These data from different sources are close enough to be integrated in a common clinical data warehouse.

Keywords: Data Warehousing; Data Management; Arthroplasty, Replacement, Hip; Equipment Safety

1. Introduction

The number of Total Hip Arthroplasty (THA) is constantly increasing in France [1]. With the reinforcement of the European regulatory constraints, there is a need to improve follow-up of patients with hip prostheses with an efficient post-marketing surveillance [2]. Complications are rare but have important consequences on the patients' quality of life (surgical site infection, deep vein thrombosis, dislocation).

[1] Corresponding Author, Thibault DHALLUIN, Department of Medical Information, University Hospital of Tours, Tours, France. Medical school, University of Tours, EA 7505 EES, Tours France 2, boulevard Tonnellé - 37044 Tours cedex 9, France, E-mail : T.DHALLUIN@chu-tours.fr

In France, there is no national cohort of hip replacement patients of sufficient quality to be reused for surveillance purposes [3]. This lack of data makes it difficult to combine clinical information on patients and technical data on devices to identify the determinants of rare and/or delayed but severe complications such as surgical site infection or luxation. Moreover, the classic manual constitution of a cohort is a long expensive process requiring a high workload for the teams. Moreover, a dynamic link between patients and outcomes could allow real-time updated surveillance [4–6].

The digitization of medical records and health examinations represents now a large re-usable data sources to monitor adverse event. These digital data could be stored following FAIR principles, in clinical data warehouses in order to provide a technical, regulatory, interoperability and security framework adapted [7,8]. The use of our data warehouses already makes it possible to track drug complications and it seems necessary to study the possibility of tracking complications after an joint replacement [9,10]. Since the 2007 regulation, the references of implanted medical devices are listed and linked to the identification of the patient for device safety purposes [11]. Our hypothesis was that the device data were comprehensive and of sufficient quality to track the different components of a hip device to monitor adverse event through a clinical data warehouse.

Our objective was to assess quantitative and qualitative concordance between claim data and device supply data in order to integrate them into a data warehouse.

2. Method

We performed a single-centre pilot cohort study between January 1, 2010 and December 31, 2019, using the clinical data warehouse of one University Hospital using a large data warehouse software eHOP [12]. We included all patients' stays of hip arthroplasty procedure or with at least one hip medical device provided. Patients younger than 18 years or opposed to the reuse of their data were excluded from the analysis.

The coverage obtained by hospital stays from matching two different sources was assessed: the surgical procedure data came from claim data (hospital discharge database PMSI), completed with a French version of the Current procedural terminology (CPT) and the medical device data came from the pharmacy supply software. The consistency of using inclusion criteria from different sources was measured by identifying the percentage of hospital stays having both a hip replacement procedure and a hip device provided.

A descriptive analysis of the main characteristics was performed: age, sex, length of stay, surgery act (replacement, repositioning, change, or reconstruction), surgery motif (arthrosis, fracture, cancer, infection, osteonecrosis or other main diagnosis) and device provided (head, stem, shell, or other) following Giori et al [6]. This descriptive analysis specifically specified the source of each of these data and missing data.

In order to know in which medical and surgical situation we were able to obtain each component, we crossed the hospital stays according to the main cause of surgery and the presence of a head, a stem and a shell device. In the same way we crossed the type of surgical procedure and the presence of a head, a stem and a shell device to evaluate the completeness according to the type of surgical procedure.

3. Results

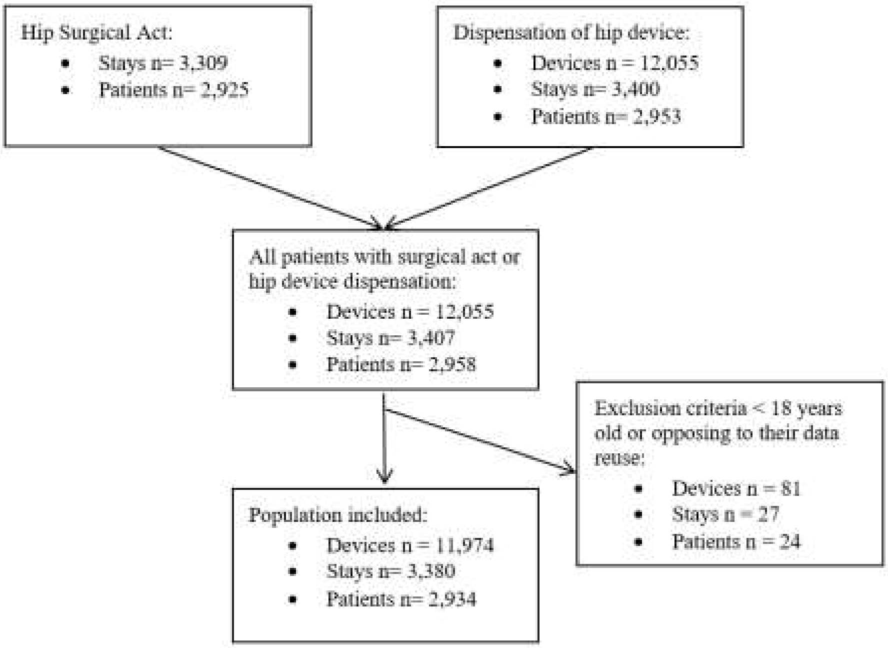


Figure 1. Flow chart

Table 1. Main characteristics of the population

Features	Descriptive Data	Source of Data
Stays (N = 3,880)		
Age (mean (sd))	71.3 (13.8)	Hospital discharge database
Sex		
- Female (%)	1,910 (43.5%)	
- Male (%)	1,470 (56.5%)	
Length of stay (mean (sd))	10.3 (10.5)	
Surgery procedure		Claim data French CPT
- Replacement (%)	2,642 (78.2%)	(Common procedure terminology)
- Change (%)	475 (14.1%)	
- Reconstruction (%)	86 (1.7%)	
- Repositioning (%)	57 (1.7%)	
- Missing value (%)	120 (3.5%)	
Cause of surgery		Claim data ICD-10
- Arthrosis (%)	1,864 (55.1%)	
- Fracture (%)	875 (25.9%)	
- Infection (%)	153 (4.5%)	
- Cancer (%)	85 (2.5%)	
- Osteonecrosis (%)	58 (1.7%)	
- Other (%)	327 (9.7%)	
- Missing value (%)	18 (<1%)	
Device (N = 11,974)		Device supply data
- Femoral Head (%)	3,444 (28.8%)	
- Femoral Stem (%)	3,055 (25.5%)	
- Acetabular Cup (%)	4,952 (41.3%)	
- Others (%)	523 (4.4%)	

Over the study period, 3,407 hospital stays corresponded to a hip replacement; of which 3,309 have a medical / surgical procedure and 3,400 have a medical device provided. Moreover, 27 stays were excluded because they were under 18 years old or opposed to the data reuse. We obtained 3,380 hospital stays, including 11,974 hip devices implanted.

The coverage between the medical / surgical procedure and the device supply data was 96.4%. Among the 3,380 hospital stays, 120 (3.6%) stays without hip surgery procedure were found.

The mean age of the patients was 71.3 years old years with 56.5% women and 43.5% men. The mean length of stay was 10.3 days. 55.1% of the stays were for hip arthrosis and 25.9% for femoral neck fracture (table 1).

Table 2. Presence of hip devices according to the main cause of surgery.

	Arthrosis	**Fracture**	**Infection**	**Cancer**	**Osteonecrosis**	**Others**
N	1,864	875	153	85	58	327
Device						
Head (%)	1,843 (98.9%)	871 (99.5%)	150 (98%)	84 (98.8%)	58 (100%)	293 (89.6%)
Stem (%)	1,801 (96.6%)	851 (97.3%)	84 (54.9%)	32 (37.6%)	56 (96.6%)	168 (51.4%)
Shell (%)	1,844 (98.9%)	864 (98.7%)	143 (93.5%)	80 (94.1%)	58 (100%)	299 (91.4%)

Table 3. Presence of hip devices according to the surgical procedure.

	Replacement	**Change**	**Reconstruction**	**Repositioning**
N	2,642	475	86	57
Device				
Head (%)	2,633 (99.7%)	428 (90.1%)	85 (98.8%)	56 (98.2%)
Stem (%)	2,608 (98.7%)	263 (55.4%)	47 (54.7%)	49 (86%)
Shell (%)	2,632 (99.6%)	425 (89.5%)	84 (97.7%)	57 (100%)

Over 90% of the hospital stays had a femoral head and shell references. The presence of femoral stems was more inconstant, especially in procedures performed for surgical site infections or cancer. Similarly, we found device references in over 98% of joint replacement procedures. In more complex procedures such as prosthesis change, reconstruction and repositioning the references were found in 50 to 100% of the cases (table 2-3).

4. Discussion

With 96.4% coverage, we obtained close data between the hospital stays obtained by medical-surgical procedures and those obtained by medical devices. The devices were found for more than 90% of the heads and shells and almost entirely for the most common clinical cases such as joint replacement in first intention for osteoarthritis.

In the case of change or repositioning surgery, the prosthesis stems were not systematically replaced and the procedure might concern only the head and the acetabulum, which might explain the procedures without stem. In the reconstructive surgery scenario, devices data included batch devices, but did not include custom prostheses, which may explain the missing stem devices in some femoral cancer reconstruction.

These results are obtained from a single centre, but these data might also be close in other centres because claims data are collected in the same way for all French healthcare facilities and device data are subject to the same traceability regulations in France. The data obtained are subject to the usual bias of information, when handling the hospital discharge database, and in the same way the device supply data the errors of information are ever possible. The number of medical records with missing device data was reasonable to be manually reviewed and corresponded to case-by-case situations, either due to data input errors or surgery where one of the pieces was actually not dispensed. The data reuse of medical device dispensed for a post-marketing surveillance and epidemiological purposes seemed possible.

The reliability of these data seemed high enough to be integrated in our clinical data warehouse. Among the next challenges, the organization of devices according to a common thesaurus seems complex considering the heterogeneous characteristics of medical devices and the lack of international common thesaurus.

References

[1] Putman S, Girier N, Girard J, Pasquier G, Migaud H, Chazard E. Épidémiologie des prothèses de hanche en France: analyse de la base nationale du PMSI de 2008 à 2014. Revue de Chirurgie Orthopédique et Traumatologique. 2017 Nov 1;103(7):S90..

[2] EUR-Lex - 32017R0745 - EN - EUR-Lex, (n.d.). https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (accessed August 3, 2021).

[3] Registre des prothèses de hanche, *SOFCOT.* (n.d.). https://www.sofcot.fr/cnp-cot/registre-des-protheses-de-hanche (accessed August 3, 2021).

[4] Resnic FS, Matheny ME. Medical devices in the real world. The New England journal of medicine. 2018 Feb 15;378(7):595-7.

[5] Dhalluin T, Fakhiri S, Bouzillé G, Herbert J, Rosset P, Cuggia M, Grammatico-Guillon L. Role of real-world digital data for orthopedic implant automated surveillance: a systematic review. Expert Review of Medical Devices. 2021 Aug 3(just-accepted).

[6] Giori NJ, Radin J, Callahan A, Fries JA, Halilaj E, Ré C, Delp SL, Shah NH, Harris AH. Assessment of Extractability and Accuracy of Electronic Health Record Data for Joint Implant Registries. JAMA network open. 2021 Mar 1;4(3):e211728-..

[7] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9.

[8] Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, Delamarre D, Raimbert V, Lemordant P, Cuggia M. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Studies in health technology and informatics. 2019 Aug 1;264:1536-7.

[9] Bouzillé G, Morival C, Westerlynck R, Lemordant P, Chazard E, Lecorre P, Busnel Y, Cuggia M. An Automated Detection System of Drug-Drug Interactions from Electronic Patient Records Using Big Data Analytics. InMedInfo 2019 Aug 21 (pp. 45-49).

[10] Sylvestre E, Bouzillé G, Chazard E, His-Mahier C, Riou C, Cuggia M. Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records. BMC medical informatics and decision making. 2018 Dec;18(1):1-8.

[11] Décret n° 2006-1497 du 29 novembre 2006 fixant les règles particulières de la matériovigilance exercée sur certains dispositifs médicaux et modifiant le code de la santé publique (Dispositions réglementaires) - Légifrance, (n.d.). https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000463293 (accessed October 1, 2021).

[12] Benchimol EI, Smeeth L, Guttmann A, Harron K, Hemkens LG, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen. 2016 Oct 1;115:33-48.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210810

FAIR Aspects of a Genomic Information Protection and Management System

Jaime DELGADO[1] and Silvia LLORENTE
Information Modeling and Processing (IMP) group – DMAG,
Computer Architecture Department (DAC),
Universitat Politècnica de Catalunya (UPC – BarcelonaTech), Spain

Abstract. To handle genomic information while supporting FAIR principles, we present GIPAMS, a modular architecture. GIPAMS provides security and privacy to manage genomic information by means of several independent services and modules that interact among them in an orchestrated way. The paper analyzes how some security and privacy aspects of the FAIRification process are covered by the GIPAMS platform.

Keywords. Genomic Information, FAIR, Protection, Access control

1. Introduction

Genomic information can be represented using different formats, depending on: kind of information stored (raw, aligned, unaligned, processed in some way, etc.), compressed or not, lossy or lossless, binarized or not, etc. In fact, the format to choose is very much related to the purpose and environment of its use.

In this context, it is also important to take into account that genomic information usually has associated metadata, which can be expressed using XML (eXtensible Markup Language), as in [1],[2], or directly in other formats [3], and can be stored inside [4], or outside the files containing the genomic information they describe (or apply to) [1],[2]. This metadata may describe the information contained in a genomic file, information about the tools or commands used in the processing pipeline, information about what is being studied (medical condition, patient, etc.) or information about security techniques used to protect the genomic information. Furthermore, it might include rules to control the access to the information [4].

Due to the specific characteristics of human genome information, which uniquely identifies a person and her relatives, it is very important to keep it safe, applying security and access control measures. But this may be difficult to achieve when few genomic information is available, since re-identifying data might be a real risk. On the other hand, Findable, Accessible, Interoperable and Reusable (FAIR) principles [5] are a desirable feature for research data, including genomic data.

Once we have the information, then we need the tools to handle it. Again, there are different approaches for this. We have designed a modular architecture, GIPAMS (Genomic Information Protection And Management System), where different

[1] Corresponding Author, Jaime Delgado, Universitat Politècnica de Catalunya (UPC - BarcelonaTech), Barcelona, Spain; e-mail: jaime.delgado@upc.edu

functionalities are provided by independent services interacting between them. As indicated in the acronym, security and privacy are key aspects in the design of the platform. Another key aspect is to provide the mentioned FAIR principles. We had previously analyzed security and privacy in the context of FAIR [6] and, in this paper, we use some of those results to validate, from a security point of view, that GIPAMS provides FAIR principles.

In the rest of the paper, we describe our platform architecture and we point out how we provide protection for genomic information while applying FAIR principles.

2. Methods - Platform Architecture

The architecture of the developed platform is an evolution of our original Multimedia Information Protection And Management System, MIPAMS [7]. As it now deals with genomic information, we have called it Genomic Information Protection And Management System, GIPAMS. [8] describes its implementation details. GIPAMS structure is depicted in Figure 1. The different modules are briefly described next:

- User Application: Access point to the whole system. It sends all requests to the Workflow Manager, which redirects to the corresponding module. An access token is required, which is provided by the Authentication Service.
- Workflow Manager: Intermediate module that acts as a unique entry point to the system to facilitate interactions with the other modules and making them transparent to the final user. Before redirecting an operation coming from the User Application, it checks if this operation is authorized using the information inside the access token.
- Authentication Service: User identification server, which uses OAuth 2.0 [9] and JSON Web Tokens [10].
- Genomic Content Service: Deals with genomic archive management, both in reading and writing operations.
- Authorization Service: Validates authorization rules. It mainly receives requests from the Workflow Manager responding to user actions, but other modules may also interact with it.
- Search Service: Performs searches over genomic information.
- Policy Service: Creates authorization rules, which are organized into policies.
- Protection Service: Creates metadata representing protection information associated to genomic information and also applies the defined mechanisms (i.e. encryption, signature, etc.) to the information.
- Report / Track Service: Deals with the reporting of operations done in the system, especially those not authorized. It helps in keeping track of illegal / unusual operations that may indicate an attempt to attack the system.
- Certification Authority: This is not a real module of the system, but something required for its proper functioning.

It is worth noting that this architecture is independent of the kind of information it handles. As indicated, we started with audiovisual information, but we also used it for other types of health information [11].

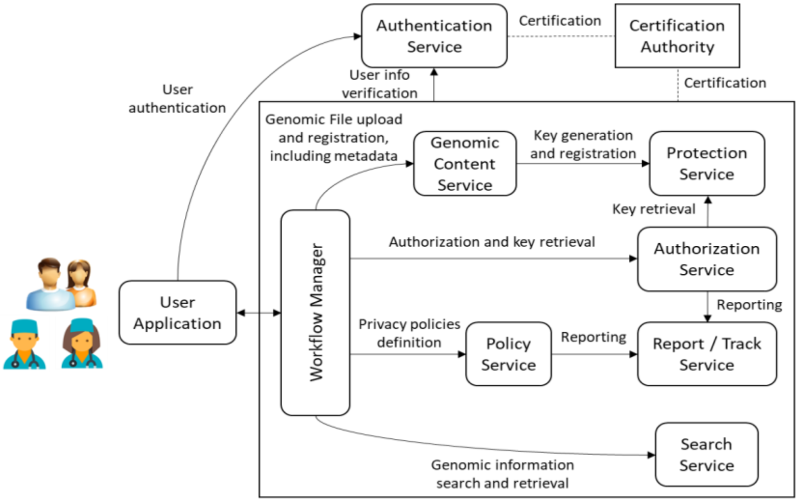


Figure 1. GIPAMS Architecture.

3. Results - FAIR principles in GIPAMS

In order to identify how GIPAMS supports the FAIR principles, we start from the ideas presented in [6], where we analyzed how FAIR principles could be applied to genomic information considering privacy and security. In particular, we discussed in detail steps 3 ("Data de-identification/anonymization") and 6 ("License attribution") of the FAIRification process described in [12]. In the work we are presenting here, however, we focus on step 6, so we consider "license attribution" as the driving concept to analyze how our system GIPAMS is FAIR without losing privacy and security.

Specifically, with respect to license attribution, in [6] we tried to answer to the following four questions: 1) How to express the licenses, 2) How to protect them and guarantee their provenance, 3) How to evaluate their authorization, and 4) How to enforce what they are controlling. The rest of the section provides a first answer to these questions in the context of GIPAMS.

3.1 Expression of licenses

The response to this first question is to use a formal language to facilitate interoperability (the I of FAIR). Rules formally expressed clearly define how to access the information (the A of FAIR). One possibility for the expression of these licenses is to use the eXtensible Access Control Markup Language (XACML) [13] and this is what we have implemented in the Policy Service. XACML allows to express the rules that control who, how and when may access specific genomic information, be it data or metadata. The expression language has an associated mechanism to evaluate the rules, based on standardized requests. In particular, the Policy Service allows the creation of rules, while their evaluation takes place in the Authorization Service.

3.2 Protection of provenance of licenses

The answer to the second question is also implemented in the Policy Service, as the policies and rules created can be protected against modification using XML Signature. Furthermore, this allows to know the origin of the license. This is an extra feature, normally not available in current standards, that we have provided for this service. From a FAIR point of view, this provides support to the A and even to the I, as before, but also, partially, helps in making information Reusable (the R), since we are confident in its origin and lack of modification.

3.3 Authorization upon licenses

As introduced in 3.1, the third question is answered in the Authorization Service, which uses the mechanisms defined in XACML [13]. "XACML Requests" including different attributes like subject, object, action or time conditions have to be defined to check if they fulfill any of the XACML rules stored in the GIPAMS' Policy Service. The request and the rule are related to an object, which can be any part of the genomic information, including metadata. Again, the A and I from FAIR are supported here.

3.4 Enforcement with licenses

The response to this last question is the GIPAMS platform itself. If the requested action is not authorized, the requested information (which is encrypted for its protection and stored in the Genomic Content Service) will not be provided. It is also possible to keep track of the actions performed in the system by means of the Reporting Module. In this context, it is worth mentioning the Search Service, that would add Findability (F) to the other three FAIR concepts. Although this is not explicit for security, it is however relevant for FAIR.

4. Discussion

We have analyzed a specific aspect of security and FAIR principles (license attribution). Solving the 4 questions raised in section 3 mainly allows to guarantee Accessibility to the genomic information, for the authorized people in the authorized circumstances. Furthermore, GIPAMS also provides Interoperability, since it uses standards for expressing and validating licenses / access rules / policies. Reusability is indirectly provided, as mentioned in 3.2. Finally, Findability is a core part of the platform.

Several GIPAMS' modules provide this FAIR support, as described in section 3, as Policy, Authorization, Genome Content, Reporting and Search modules. In any case, it is the complete platform who supports the license attribution features.

Another important aspect of a system like GIPAMS is that we might have other "xIPAMS" platforms by providing specific Content services. In other words, our architecture is independent of the kind of content, since we might include different specific Content services. In GIPAMS, we have a Genomic Content service, while in MIPAMS we have a Multimedia Content service. As indicated at the end of section 2, the architecture could be used for other types of health information, as we described in [11]. This means that the provision of security and privacy on one hand, and FAIR principles on the other, is not only valid for genomic information, but also for other eHealth information. This is also very useful when trying to integrate genomic information with current health records.

5. Conclusions

A modular and distributed approach for the management of genomic information facilitates following the FAIR principles. This is accomplished with our Genomic Information Protection And Management System (GIPAMS). We have analyzed how GIPAMS supports the FAIR (Findable, Accessible, Interoperable and Reusable) principles from a security and privacy point of view. We have started from previous work and reached the expected conclusions. Our focus has been on licenses to control the access to information. Some identified GIPAMS' modules, and the complete platform in general, mainly support the Accessibility and Interoperability FAIR principles, but we may also consider the other two.

On the other hand, it is worth noting that, although GIPAMS is intended for the handling of genomic information, other "xIPAMS" platforms may provide services over other eHealth content. Some of our future work concentrates in designing and developing different xIPAMS platforms that would support integration of genomic and other health information, guaranteeing security and privacy and supporting all FAIR principles. We also plan to apply this to a real clinical environment in a new project in Spain.

Acknowledgements

This work is partly supported by the Spanish Government (GenClinLab-Sec, PID2020-114394RB-C31) and by the Generalitat de Catalunya (2017 SGR 1749).

References

[1] National Center for Biotechnology Information (NCBI). 2021. https://www.ncbi.nlm.nih.gov

[2] European Nucleotide Archive (ENA). 2021. https://www.ebi.ac.uk/ena/browser/about

[3] Sequence Alignment / Map (SAM) Format Specification. 2018. https://samtools.github.io/hts-specs

[4] ISO/IEC 23092-3:2020, Genomic information representation - Part 3: Metadata and application programming interfaces (APIs). 2020. https://www.iso.org/standard/75625.html

[5] Wilkinson, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. 2016. https://doi.org/10.1038/sdata.2016.18

[6] Delgado J, Llorente S. Security and Privacy when Applying FAIR Principles to Genomic Information. Stud Health Technol Inform. 2020 Nov 23;275:37-41. doi: 10.3233/SHTI200690. PMID: 33227736.

[7] Llorente S, Rodriguez E, Delgado J, Torres-Padrosa V. Standards-based architectures for content management. IEEE MultiMedia. 2012 Nov 29;20(4):62-72.

[8] Delgado, J.; Llorente, S; Reig, G. Implementation of privacy and security for a genomic information system. pHealth 2021. 2021.

[9] IETF. The OAuth 2.0 Authorization Framework, 2012. https://datatracker.ietf.org/doc/html/rfc6749

[10] IETF. JSON Web Token (JWT), 2015. https://datatracker.ietf.org/doc/html/rfc7519

[11] Delgado J, Llorente S. Privacy provision in eHealth using external services. Stud Health Technol Inform. 2015;210:823-7. PMID: 25991269.

[12] GO FAIR, FAIRification process, 2021. https://www.go-fair.org/fair-principles/fairification-process

[13] OASIS, eXtensible Access Control Markup Language (XACML) v3.0, 2017. http://www.oasis-open.org/specs/index.php#xacmlv3.0

doi:10.3233/SHTI210811

Extraction of Temporal Structures for Clinical Events in Unlabeled Free-Text Electronic Health Records in Russian

Anastasia A. FUNKNER[a,1] Dmitrii A. ZHURMAN[a] and Sergey V. KOVALCHUK[a,b]
[a] *ITMO University, Saint Petersburg, Russia*
[b] *National Almazov Medical Research Centre, Saint Petersburg, Russia*

Abstract. The important information about a patient is often stored in a free-form text to describe the events in the patient's medical history. In this work, we propose and evaluate a hybrid approach based on rules and syntactical analysis to normalise temporal expressions and assess uncertainty depending on the remoteness of the event. A dataset of 500 sentences was manually labelled to measure the accuracy. On this dataset, the accuracy of extracting temporal expressions is 95,5%, and the accuracy of normalization is 94%. The event extraction accuracy is 74.80%. The essential advantage of this work is the implementation of the considered approach for the non-English language where NLP tools are limited.

Keywords. time expression extraction; normalization; syntactical parsing; corpus; clinical text mining; machine learning

Extraction of temporal expressions from electronic health records (EHR) helps restore the chronology of the patient's diseases and order all his/her events on a timeline. Extracted temporal expressions and their events make data more findable, interoperable, and reusable according to FAIR principles [1]. There have been four competitions for the extraction of temporary structures in clinical texts [2]. However, most methods and approaches are not applicable for clinical texts in Russian because of the lack of labelled corpora [3]. Previously we developed an unsupervised approach to extract sentences with explicit temporal expressions but that approach has its drawbacks: imprecise retrievable constructions and difficulty in assessing obtained results [4].

Firstly, it is necessary to implement the extraction of temporal expressions (TEs) from sentences using rule-based methods. These TEs should be normalized to a single format (YYYY-MM-DD). Secondly, sentences with TEs should be parsed the syntactical parsers. Thirdly, we need to find a path from the defined TE to the right event in the syntactic tree. The algorithm to extract events is shown in Figure 1a. We compare common syntactic parsers for Russian and choose DeepPavlov because of its regular updates and detailed documentation. In this implementation, we use Spacy for writing rules because it shows a higher processing speed (35 sentences per sec upon 7 sentences per sec in Yargy). We develop 260 rules for TEs detection in Russian. For normalization, we used ready-made Python libraries dateparser and rutimeparser.

[1] Corresponding Author, ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia; E-mail: funkner.anastasia@itmo.ru.

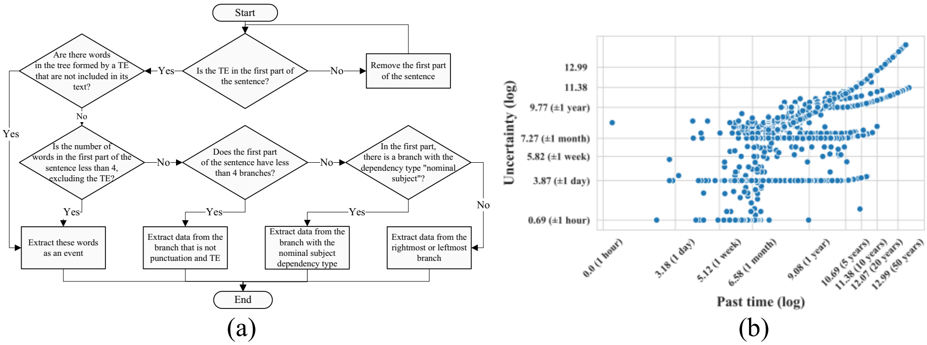

Start
Are there words in the tree formed by a TE that are not included in its text?
Yes
Is the TE in the first part of the sentence?
No
Remove the first part of the sentence
No
Yes
Is the number of words in the first part of the sentence less than 4, excluding the TE?
No
Does the first part of the sentence have less than 4 branches?
No
In the first part, there is a branch with the dependency type "nominal subject"?
No
Yes
Yes
Yes
Extract these words as an event
Extract data from the branch that is not punctuation and TE
Extract data from the branch with the nominal subject dependency type
Extract data from the rightmost or leftmost branch
End
(a)
Uncertainty (log)
12.99
11.38
9.77 (±1 year)
7.27 (±1 month)
5.82 (±1 week)
3.87 (±1 day)
0.69 (±1 hour)
0.0 (1 hour)
3.18 (1 day)
5.12 (1 week)
6.58 (1 month)
9.08 (1 year)
10.69 (5 years)
11.38 (10 years)
12.07 (20 years)
12.99 (50 years)
Past time (log)
(b)

Figure 1. Methods and results: (a) the algorithm for events extraction from a syntactic tree, and (b) logarithmic uncertainty for retrieved events.

The Cardiology Research Institute (Tomsk, Russia) provides anonymized EHRs, consisting of events and patient information. The Research Institute dataset includes 7777 sentences with temporal expressions (7277 and 500 for train and test sets). On the manually labelled test dataset, the accuracy of extracting TEs is 95.5%, the accuracy of normalization is 94%, and the events extraction accuracy is 74.8%.

We apply a trapezoidal membership function with a remoteness parameter to assess the uncertainty of extracted events. Figure 1(b) shows the uncertainty (log-log scale) for 6344 events. As can be seen, events of one category (known day, only month or year) form distinct line patterns. These lines bend as the age ratio increases linearly. Event uncertainty shows how reliable can be extracted events. Uncertainty scores can be used to build more accurate models as an additional feature.

In this paper, we propose an approach for extracting temporal structures and events in the absence of labelled data corpora for medical texts. With proper syntactic parsers, the models can be implemented for any other language. As the approach is focused on working without a labelled corpus, we believe it could find broader application in other languages with a lack of available public corpora and NLP tools in the medical domain.

Acknowledgements.

This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:1–9.

[2] Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Informatics Assoc 2013;20:828–35.

[3] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. J Biomed Semantics 2018;9:1–13.

[4] Funkner AA, Kovalchuk S V. Time Expressions Identification Without Human-Labeled Corpus for Clinical Text Mining in Russian. Comput. Sci. -- ICCS 2020, Springer International Publishing; 2020, p. 591–602.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210812

One Digital Health Is FAIR

Arriel BENIS[a,1] and Oscar TAMBURIS[b]
[a] *Faculty of Industrial Engineering and Technology Management, Holon Institute of Technology, Holon, Israel*
[b] *Department of Veterinary Medicine and Animal Productions, University "Federico II", Naples, Italy*

Abstract. The One Digital Health framework aims at transforming future health ecosystems and guiding the implementation of a digital technologies-based systemic approach to caring for humans' and animals' health in a managed surrounding environment. To integrate and to use the data generated by the ODH data sources, "FAIRness" stands as a prerequisite for proper data management and stewardship.

Keywords. One Health, Digital Health, One Digital Health, FAIR, FAIRness

1. Introduction

The One Digital Health (ODH) [1] framework allows the analysis of the digital health ecosystem components through different perspectives focusing on how technologies may support healthcare and well-being activities. An ODH intervention can support the management of such a web (i.e., human, animal, and environmental) of digital interconnections. To integrate and to use the data generated by the ODH data sources, "FAIRness" stands as a prerequisite for proper data management and stewardship [2,3]: the FAIR Principles provide guidelines for the publication of those digital resources (or digitalities) whose combination and implementations set up the shape of an ODH intervention, for making them Findable, Accessible, Interoperable, and Reusable [2,3]. Besides the existing FAIRness metrics, new ones need therefore to be developed within new contexts such as ODH. How does the ODH framework support FAIR?

2. Materials and Methods

The ODH "Steering Wheel" is built around 2 keys (One Health, Digital Health), 3 perspectives (individual health and well-being, population and society, ecosystem), and 5 dimensions (citizens' engagement, education, environment, human and veterinary health care, Healthcare Industry 4.0) [1]. The digital technologies (digital-ities) to be singled out and analyzed within the ODH dimensions prism aim to: increase animal welfare and account for "how" humans affect animals' lives, health, and interactions; relate to how technology has been embedded in human experiences and activities; aim to support the management and governance of the complex interactions between humans,

[1] Corresponding author, Dr. Arriel Benis, Faculty of Industrial Engineering and Technology Management, Holon Institute of Technology, Golomb St 52, Holon, 5810201, Israel; e-mail: arrielb@hit.ac.il.

animals, and their ecosystems. FAIR spotlights the capacity of computational systems to find, access, interoperate and reuse data with a minimum of human interventions, due to the increasing volume, velocity, and variability of data. This means that, to get to an optimal ODH-ness (i.e. an effective supply of an ODH intervention), each investigated digitality involved is globally evaluated, also in terms of its FAIRness with adapted metrics.

3. Results

The developing ODH-ness compliance analysis assessment comprises a FAIRness evaluation management component. The design and deployment of an ODH intervention imply for data to be: Findable because the digitalities involved are part of the study and collection of all the data related to the interconnection between systems' needs; Accessible via standardized protocols, to leverage the available common substrates of data, information, and knowledge stemming from digital biodiversity; Interoperable as a consequence of the awareness to establish an ecosystem capable of seamless, secure health data exchange and processing, to deal with the shared risks between animal and human populations; Reusable to allow a systematic, continuous, and intelligent integration of big, smart, and multidimensional data to be exchanged by the digitalities involved.

4. Discussion and Conclusion

Despite the efforts performed to deploy Open Science data stewardship, the comprehensive view that the ODH framework can provide is still lacking. ODH requires adopting new kinds of data environments, technologies, and standards. The COVID-19 pandemic has forced people to dwell on the close relationships among the environment, animals, and humans. Initiatives as the FAIR4Health project encouraged the health research community to FAIRify, share, and reuse their datasets derived from publicly-funded research initiatives [4]. The ODH is FAIR. ODH pillars are the availability (findability, accessibility) of human, animal, and environmental data allowing a unified understanding of complex interactions (interoperability) over time (reusability). It is therefore a prolific landscape that joins FAIR for global health as an interdisciplinary and unifying layer by developing "fair" ODH interventions.

References

[1] Benis A, Tamburis O, Chronaki C, Moen A. One Digital Health: A Unified Framework for Future Health Ecosystems. J. Med. Internet Res. 2021; 23: e22189.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3: 1–9.

[3] Wilkinson MD, Sansone S-A, Schultes E, et al. A design framework and exemplar metrics for FAIRness. Scientific Data. 2018; 5: 1–4.

[4] FAIR4Health, https://www.fair4health.eu/en/project# (Accessed on September 9th, 2021).

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210813

Investigating the FAIR Equivalency in National Guidance in Health in Kenya

Esther Thea INAU[a,1], Reginald NALUGALA[b,c], William Muhadi NANDWA[d], Fredrick OBWANDA[c], Antony WACHIRA[e], Dagmar WALTEMATH[a], Antonio CARTAXO[f] and Atinkut ZELEKE[a]
[a] *Department of Medical Informatics, University Medicine Greifswald, Germany*
[b] *Institute of Social Transformation, Tangaza University, Kenya*
[c] *VODAN-Africa, Kenya*
[d] *Pumwani Hospital, Kenya*
[e] *Strathmore University, Kenya*
[f] *The Globalisation, Accessibility, Innovation & Care Network, Leiden University, The Netherlands*

Keywords. Data stewardship, FAIR, COVID-19, Kenya

1. Introduction

The growth in the public health sector in Sub-Saharan Africa has been supported by the implementation of various Health Information Systems [1]. The implementation of the FAIR data principles is advocated as an important cornerstone in research data management [2]. However, the implementation requires a comprehensive understanding of the already existing infrastructure and the demands of the implementing communities [3]. The Virus Outbreak Data Network (VODAN) project aims to integrate FAIRified data on the SARS CoV-2 virus in Sub-Saharan Africa (SSA) countries [4,5]. FAIR data principles are better implemented in Europe than in Africa, where there is hardly implementation at all [4]. Undefined data ownership in Africa is among the obstacles to comprehensive data stewardship during the COVID-19 pandemic [6]. This work explores the documented guidance authored by the governmental authorities in Kenya from 2006 to 2019 to direct the health and ICT sectors. We review the existent background regarding the policies, Acts, national strategies and national guidelines that may influence the uptake of the FAIR data principles in Kenya and further enable a FAIR digital data health infrastructure in Africa for reporting and research. The results serve to inform on the feasibility of FAIR implementaion within a framework of national relevance.

2. Methods

We conducted a qualitative cross-sectional study on 14 documents authored by the national authorities in Kenya from 2006 to 2019 to direct the health and ICT sectors.

[1] Corresponding Author, Esther Inau; E-mail: inaue@uni-greifswald.de.

Here we measure the convergence between the FAIR data principles and the existing regulatory frameworks in Kenya's health data stewardship sector. We examined the document collection with respect to explicit mentions of the FAIR data principles. If no mention of FAIR had been found, the documents were further examined to determine direct mentions of the 15 FAIR data facets or of concepts representing them [4]. Our investigation is based on the "FAIR Equivalency" index, which indicates the degree of agreement between Kenya's national regulatory situation and the FAIR principles [6].

3. Results and Discussion

Our analysis shows that the FAIR data principles are not explicitly mentioned, but the underlying equivalent concepts are indeed covered. The overall FAIR equivalence score is 43,79 % (Table 1). The scores per document show a great variation from 0 (0%) to 14 (93.3%). The analysis shows that the leadership is yet to make any provisions for the introduction and implementation of the FAIR data principles. However, the need for interoperability among heterogenous systems, has been comprehensively described.

Table 1. FAIR equivalence score for 14 documents with respect to the 15 facets of the FAIR data principles

FAIR data Principle (n=15)	Expected EQ max score	FAIR EQ score (%)
Findable (4)	56	27 (48.2)
Accessible (4)	56	27 (48.2)
Interoperable (3)	42	20 (47.6)
Reusable (4)	56	18 (28.52)
Total score	**210**	**92 (43.79)**

4. Conclusion

Our evaluation reveals that there is no explicit uptake of the FAIR data principles in the health domain in Kenya. However, the equivalent of the FAIR concepts exists under a different name. We recommend that the leadership be offered a detailed introduction to the FAIR data principles and the steps necessary to FAIRify health data.

References

[1] Njeru I, Kareko D, Kisangau N, Langat D, Liku N, Owiso G, et al. Use of technology for public health surveillance reporting: opportunities, challenges and lessons learnt from Kenya. BMC Public Health. 2020;20(1):1101.

[2] Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? European journal of human genetics. 2018;26(7):931-6.

[3] Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence. 2020;2(1-2):10-29.

[4] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.

[5] Mons B. The VODAN IN: support of a FAIR-based infrastructure for COVID-19. European Journal of Human Genetics. 2020;28(6):724-7.

[6] van Reisen M, Oladipo F, Stokmans M, Mpezamihgo M, Folorunso S, Schultes E, et al. Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced Genetics. 2021;2(2):e10050.

Section III

FAIR and COVID-19 (and Other Infective Diseases) Research Data

This page intentionally left blank

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210815

Transfer of Clinical Drug Data to a Research Infrastructure on OMOP – A FAIR Concept

Ines REINECKE[a,1], Michéle ZOCH[a], Markus WILHELM[a], Martin SEDLMAYR[a] and Franziska BATHELT[a]
[a] *Institute for Medical Informatics and Biometry at Carl Gustav Carus Faculty of Medicine at Technische Universität Dresden, Germany*

Abstract. Generating evidence based on real-world data is gaining importance in research not least since the COVID-19 pandemic. The Common Data Model of Observational Medical Outcomes Partnership (OMOP) is a research infrastructure that implements FAIR principles. Although the transfer of German claim data to OMOP is already implemented, drug data is an open issue. This paper provides a concept to prepare electronic health record (EHR) drug data for the transfer to OMOP based on requirements analysis and descriptive statistics for profiling EHR data developed by an interdisciplinary team and also covers data quality issues. The concept not only ensures FAIR principles for research, but provides the foundation for German drug data to OMOP transfer.

Keywords. EHR, data quality, drug administration, OHDSI, OMOP, FAIR

1. Introduction

The pandemic of coronavirus disease 2019 (COVID-19) has shown the need of standardized and reproducible research data, especially regarding drug administration, as observational studies are important to gain evidence, learn on real-word data and improve the COVID-19 patient treatment and their effects in the future [1]. However, those studies highly depend on the level of data quality, interoperability and reproducibility, even more if they are proceeded in a multi-centric environment [2].

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) developed by the Observational Health Data Sciences and Informatics (OHDSI) is one option to foster reliability of retrospective, observational studies on real-world data [3] and compared to other CDMs, e.g. i2b2, PROCNet it best supports those studies [4]. OMOP comes with standardized vocabularies and terminologies, methods for data analyses and data quality checks while ensuring FAIR principles [5,6]. However, the main prerequisite to take advantage of OMOP and corresponding tools is the storage of patient electronic healthcare record (EHR) data in an OMOP conformed way. Although this is already tackled (for a nationally consolidated core data set) by the MIRACUM project [7] of the German Medical Informatics Initiative, drug data is still an open issue. Mainly because of the fact that drug administration is often documented in a non-

[1] Corresponding Author, Ines Reinecke, Institute for Medical Informatics and Biometry at Carl Gustav Carus Faculty of Medicine at Technische Universität Dresden, Germany; E-mail: ines.reinecke@tu-dresden.de

standardized way and using unstructured data. Therefore this paper aims to provide a concept on data preparation for EHR drug data which is documented during in-patient visits at a German university hospital to be used in the OMOP for research in order to increase FAIR data principles for observational research on real-world data.

2. Methods

To move from heterogenous and proprietary EHR data to OMOP that aligns to the FAIR principles [5], we developed a target oriented concept based on medical expertise and an EHR as well as an OMOP analysis (Figure 1). Working with EHR data in research requires a deep understanding of the original data (e.g data origin, data completeness, data correctness, data structure) and the given target environment for research [2]. Thus we built a multidisciplinary team of data and computer scientists as well as pharmacists.

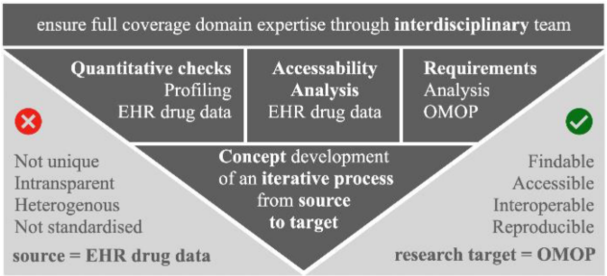


Figure 1. Methodology to derive a concept

A **requirements analysis** was done to determine the data elements and terminologies necessary to work with drugs in OMOP and to run OHDSI network studies. The analysis was done by [1] identifying relevant OMOP tables and data elements and [2] reviewing existing OHDSI network studies. Task (1) was done based on OMOP version 5.3.1 since this is the latest version supported by the current available OHDSI software stack. Task (2) was done for 29 OHDSI network studies [8] identified by Reinecke [9] in a Scoping Review. We checked the study protocols for those publications and determined if drug data was relevant and whether the drug dose information was required to answer the research question. The **accessibility** of EHR drug administration data in the corresponding IT system was analyzed in general and in comparison to the identified requirements. Access to drug documentation for intensive care was restricted and therefore excluded, additionally this study is limited to drug products registered by the German Federal Institute for Drugs and Medical Devices (BfArM) in a drug catalogue. A data profiling by **quantitative checks** of the drug prescription data element *dose unit* was done for all in-patient cases (approx. 55000) in the year 2020, for example we determined the quantitative ratio between free-text usage and drug catalogue reference in the drug administration data element *drug name*. Additionally a quantitative analysis of the *dose unit* values of the drug prescription data was done. Based on the results of the above methods a **concept of an iterative process** to convert EHR drug data into the OMOP was developed, which includes repetitive discussions with pharmacists (see Figure 3).

3. Results

The requirements analysis was limited to the OMOP *drug_exposure* table. The OMOP *drug_strength* table does not contain clinical data but drug concept information with dose and unit of drug ingredients, components and procucts that supports the standardization for drug utilization analysis. Table 1 shows a minimum list of required data elements in the OMOP table *drug_exposure*. The OMOP table column *quantity* was included, although it is not required by the CDM conventions, but we identified this information as required to conduct studies that need the drug dose information to answer the research question.

Table 1. Minimum list of required data elements in the OMOP table drug_exposure

drug_exposure	**description**	**EHR availability**
drug_exposure_id	unique key in the table	yes
person_id	reference to the patient identifier in the person table	yes
drug_concept_id	standard concept of domain drug	only drug names (catalogue/free-text)
drug_exposure_start_date	determines the start date of a drug exposure	yes
drug_exposure_end_date	determines the end date of a drug exposure	yes
drug_type_concept_id	specifies the type of a drug concept (e.g. EHR medication list, EHR prescription)	yes
quantity	based on dose form it refers to - The amount of tablets for clinical drugs with a fixed dose form - The amount of ingredient for divisible, liquid dose forms like injections	yes

The review of the studies identified 23 of 29 studies requiring drug data. Most of them used drug data based on RxNorm ingredient level, with no drug dose information. Only 2 studies were taking dose information into account for research. The analysis of the EHR drug data determined the availability of the identified and required data elements in the EHR system as shown in Table 1. The drug_concept_id is not available in the EHR system. Rather the drug name exists in the EHR system either as drug catalogue entry identifier or free-text information. The BfArM drug catalogue includes ingredient with ATC codes of active ingredients for each catalogue entry. Figure 2 visualizes the distribution of drug catalogue entries and free-text information. 59.50% of the drug prescription have a reference to the catalogue with ATC code and dose information. The other 40.50% have free-text only. Figure 3 shows the iterative process to prepare the EHR drug data to facilitate research based on OMOP. The data clean-up has to be done for free-text drug data by an appropriate domain expert. First computer scientists develop algorithm to extract ATC codes, dose and unit information from the free-text. Second the results get evaluated by pharmacists. The data clean-up is not needed for EHR data comprising drug catalogue entries. The concept mapping consists of step 3 and 4, where first the ATC concept gets mapped to the RxNorm ingredient concepts and second the RxNorm ingredient concept and the dose information has to be mapped to the RxNorm drug component. In step 5 the prepared drug data is moved to the OMOP database by an ETL job. Finally a data quality assessment using the OHDSI data quality dashboard [10] will be done. The results of the data quality assessment will be used to improve the previous steps if needed in the next iteration.

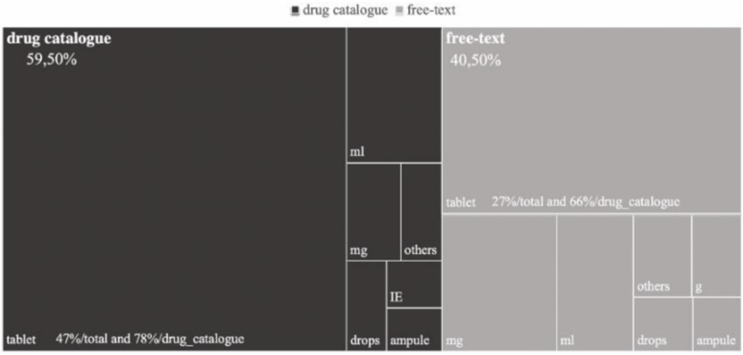


Figure 2. Availability of drug data in the EHR system

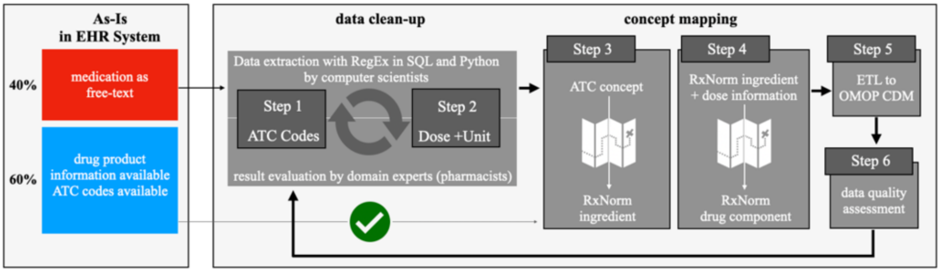


Figure 3. Concept drug data transfer from EHR system to OMOP, ensuring data quality (Kahn et. al.)

4. Discussion

The concept for transforming real-world drug data to OMOP works towards the standardization and interoperability of data as well as the reproducibility of studies while ensuring data quality. In particular the extraction of ATC codes, dose and unit information from free-text is promising regarding the step towards applying FAIR principles for in-patient care data. With the continuous result evaluation by domain experts we ensure correctness of the original data and its meaning. The steps from ATC concept to RxNorm drug components further increase the standardization and thus the semantic interoperability. The subsequent implementation of the ETL processes enables research based on OMOP. Although ETL processes on German EHR data already exist [11], our approach extends the limitation on claim data to support drug data and thus is crucial to participate in observational research on OMOP in the future. It is essential for the introduction of the OHDSI Data Quality Dashboard to check for completeness, conformity and plausibility [12] of data in OMOP. Our concept is limited to EHR drug data prescription during a hospital stay. Integrating Intensive care unit (ICU) data needs further investigation since drugs get often applied continuously with a changing dose rates over time that rises new challenges in terms of dose calculations and conversion to OMOP. Drug history data is often part of free-text medical history records and requires implementation of NLP algorithm. In a first proof of concept the data clean-up has been applied to a small data set for a very specific clinical research question. In a next step, the concept will be applied systematically to all available EHR drug data for evaluation and quality assessment and to provide feedback on overall EHR drug data quality back to the patient care teams.

5. Conclusions

This paper provides a concept that closes the gap between EHR drug data and the requirements given by the common data model OMOP that focuses on improving the FAIRness of real-world data for research. It builds the foundation on converting German drug data to international standardized research environments and is an important step to enable German research groups to participate in studies in the OHDSI community.

Declarations

Acknowledgement: This work is part of the project MIRACUM, funded by the German Ministry of Education and Research (FKZ 01ZZ1801A/L)

Author contributions: IR: conception of work, data analysis, concept development, results generation; All authors contributed ideas to the study and manuscript editing/revising. All authors approved the submitted manuscript and take responsibility for its scientific integrity. This study was performed by IR to (partially) fulfill the requirements for obtaining the academic degree "Dr. rer. medic." from the Technische Universität Dresden.

References

[1] EMA sets up infrastructure for real-world monitoring of treatments and vaccines [Internet]. [cited 2021 Jul 30]. Available from: https://www.ema.europa.eu/en/news/covid-19-ema-sets-infrastructure-real-world-monitoring-treatments-vaccines

[2] Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021 Mar 2;23(3):e22219.

[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.

[4] G Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Dec;64:333-341.

[5] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Dec;3(1):160018.

[6] OHDSI. The book of OHDSI, Chapter 3.7 FAIR Guiding Principles [Internet]. [cited 2021 Jul 30. Available from: https://ohdsi.github.io/TheBookOfOhdsi/OpenScience.html#ohdsi-and-the-fair-guiding-principles

[7] Prokosch H-U, Acker T, Bernarding J, Binder H, Boeker M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med. 2018 Jul;57(S 01):e82–91.

[8] Reinecke I. literature list of OHDSI studies [Internet]. Zenodo; 2021 [cited 2021 Jul 29]. Available from: https://zenodo.org/record/5145048

[9] Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP – A Scoping Review. Stud Health Technol Inform. Forthcoming 2021

[10] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality [Internet]. Health Informatics; 2021 Mar [cited 2021 Jul 29]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.03.25.21254341

[11] Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform. 2018 Jan;9(1):54–61.

[12] Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs. 2016 Sep 11;4(1):18.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210816

Beyond the FAIRness of COVID-19 Data: What about Quality?

Fabrizio PECORARO[1] and Daniela LUZI
Institute for Research on Population and Social Policies, National Research Council, Rome, Italy

Abstract. Different datasets have been deployed at national level to share data on COVID-19 already at the beginning of the epidemic spread in early 2020. They distribute daily updated information aggregated at local, gender and age levels. To facilitate the reuse of such data, FAIR principles should be applied to optimally find, access, understand and exchange them, to define intra- and inter-country analyses for different purposes, such as statistical. However, another aspect to be considered when analyzing these datasets is data quality. In this paper we link these two perspectives to analyze to what extent datasets published by national institutions to monitor diffusion of COVID-19 are reusable for scientific purposes, such as tracing the spread of the virus.

Keywords. FAIR, data quality, COVID-19, institutional datasets, data reusability

1. Introduction

Already from the beginning of the COVID-19 pandemic in March 2020 national and international authorities started to develop and update datasets to provide data to researchers, journalists, health care providers as well as public opinion. This data became one of the most important sources of information, commonly daily updated, to be analysed by scientists to investigate this epidemic period. Data is examined by the research community not only to monitor the COVID-19 diffusion across countries and localities for research purposes, but also to gain insights and propose better containment measures and policies. To facilitate the comparability and reuse of this data, one of the first target is to make these datasets compliant with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [1]. These principles are gaining consensus within scientific communities, with different initiatives carried out in the healthcare domain [2] at national and international level with the aim of promoting their adoption and implementation when defining and sharing research data. Despite compliance with the FAIR principles is mainly met to research results, such as clinical trials or human genomics, in this paper we pose the attention on datasets published by national institutions to report aggregate data on the diffusion of COVID-19. Furthermore, even if the compliance with the FAIR principles may be considered as a proxy for data quality assessment, they do not, in themselves, cover the crucial aspects of intrinsic data quality. However, to establish credibility

[1] Corresponding Author, Fabrizio Pecoraro, IRPPS-CNR, via Palestro 32, 00185 Rome, Italy; E-mail: f.pecoraro@irpps.cnr.it

studies that use healthcare data are increasingly expected to demonstrate that the quality of data is adequate to support research conclusions [3]. This is particularly true considering COVID-19 surveillance data that represents an essential tool to monitor trends in the epidemics, to conduct risk assessments and to timely guide preparedness and response measures [4]. For these reasons, aim of this paper is to capture the level of FAIRness of the above-mentioned institutional datasets also under the lens of the data quality model proposed by the ISO 25012 [5] which is used to define data quality requirements guiding software development.

2. Materials and Methods

COVID-19 institutional datasets available at national level in six European countries (Belgium, France, Germany, Italy and UK) were included in the analysis. They were identified carrying out a literature review in the LitCovid [6] portal that tracks COVID-19 related articles in PubMed. In particular, we concentrated on the Epidemic Forecasting section to identify datasets adopted to model the spread of COVID-19 focusing on at least one of the above-mentioned countries. Data availability statement of each paper has been analysed to extract the source of information applied to perform the analysis. Results of this review are updated at the end of June 2021.

The extracted datasets have been firstly analysed under the FAIRness perspective checking their compliance to the 15 sub-principles reported in [1]. Considering data quality, different assessment methods and models have been proposed in the literature [7] most of them defined in specific health context (e.g. prevention) or focusing on a specific disease (e.g. cancer). This perspective differentiation led authors to adopt different data quality characteristics depending on relevant points of view. In this paper we adopted the data quality model reported in ISO 25012 [5] which is widely used in different domains both at industrial and scientific levels. This standard is based on 15 characteristics classified into two categories: 1) *inherent data quality* that refers to the degree to which data quality characteristics have intrinsic potential to satisfy implicit data needs and 2) *system-dependent data quality* that refers to the degree to which data quality is achieved and preserved through an information system and is dependent on the specific technological context in which the data is used. In this paper we focus the attention on the inherent data quality characteristics.

3. Results

3.1. Analysis of national datasets on COVID-19

Among the 1700 papers published within the Epidemic Forecasting section of the LitCovid platform, 338 reported information on at least one of the six countries involved in this analysis. Almost three-quarters of them (N = 256) were excluded from the analysis as they are based on datasets published by international bodies (e.g. WHO) or adopted data collected specific studies (e.g. surveys, hospital). Table 1 shows the list of datasets adopted in the 82 remaining papers which also makes references to the institutions that curate them.

Table 1. Source of institutional datasets reported at national level

Country	Publisher	Source / Dataset
Belgium	Sciensano	https://hepistat.wiv-isp.be/Covid/
France	Public Health System	https://www.data.gouv.fr/fr/pages/donnees-coronavirus
Germany	Robert Koch Institute	https://npgeo-corona-npgeo-de.hub.arcgis.com/ https://github.com/jgehrcke/covid-19-germany-gae
Italy	Civil Protection Department	https://github.com/pcm-dpc/COVID-19
Spain	Carlos III Health Institute	https://cnecovid.isciii.es/covid19/; https://github.com/datadista/datasets/tree/master/COVID%2019
UK	Public Health England	https://coronavirus.data.gov.uk/

3.2. Analysis of FAIR principles

Table 2 shows the level of compliance of each dataset to the main FAIR principles. Considering the presentation of data, all countries defined a specific section of the institutional website to describe which data are exposed. Among them, Italy, Germany and Spain adopt the GitHub service that allows the download of CSV and JSON files directly or through the adoption of the GitHub REST API. Similarly, UK and Germany provide data with self-developed API that can also be used to download data in CSV or JSON formats. This presentation of data not only simplify the accessibility of datasets, but also ensures their findability given the permanent link through which researchers can access data routinely. Conversely, data on France and Belgium can be accessed only by downloading CSV files reported in the relevant web pages. In this case the unique identifier as well as its stability is not easily verifiable.

Table 2. Assessment of the FAIR principles in each national institutional dataset

	Belgium	France	Germany	Italy	Spain	UK
			Findable			
F1. Unique ID	HTML	HTML	API	GitHub	GitHub	API
F2 & F3. Metadata richness & ID	Limited in PDF (English)	Limited in CSV (English)	Limited in Web pages (German)	Limited in Web pages (English)	Limited in Web pages (English)	Limited in Web pages
F4. Metadata	No	No	No	No	No	No
			Accessible			
A1. Retrievability	File	File	API	API	API	API
A1.1. Protocol	CSV	CSV	API	Github	Github	API
A1.2. Auth	N/A	N/A	N/A	N/A	N/A	N/A
A2. Metadata	N/A	N/A	N/A	N/A	N/A	N/A
			Interoperable			
I1. Language	No	No	No	No	No	No
I3. Vocabulary	No	No	No	No	No	No
I4. Reference	No	No	No	No	No	No
			Reusable			
R1. Accurate	No	No	No	No	No	No
R1.1. License	Open data	Open data	Open data	Open data	Open data	Open data
R1.2. Origin	Not clear	Not clear	Not clear	Partial	Not clear	Not clear
R1.3. Standard	No	No	No	No	No	No

Considering metadata all countries provide a limited set of descriptive information, such as description and data type, along with examples describing them. Moreover, in all countries the association between a metadata file and the dataset is not explicit or even not reported. In particular, Belgium reports a codebook in a PDF file written in

English, while Germany, Italy, Spain and UK report metadata and description of indicators in specific web pages of the dataset website. France is the only country that provides a set of CSV files associated with each CSV data file reporting metadata and information about relevant indicators. However, the association between data and metadata files is not straightforward with no cross references in the documentation. All countries provide access to both data and metadata with no authentication or authorization procedures needed.

Looking at the interoperability principles, the absence of controlled vocabularies, ontologies, thesauri as well as of a data model make the integration of data and the performance of a cross-country analysis hard to be accomplished. Moreover, even if all countries, except Germany, report the description of indicators also in English, variables are generally instantiated using the original language considering both the name and the value of the indicator. The reuse of data for statistical purposes is also affected by the absence of a detailed description of the workflow that led to the collection of data. In particular, data flow and provenance of data are not sufficiently reported in each website, this is mainly critical in regional-based countries where information are daily transmitted by each region to national authorities. Lack of standardized collection of data have been reported in Italy [8] as well in Spain [9] where, each regions might count case numbers and tests with different criteria. Within the reuse of data all countries release data under the Creative Commons rules.

3.3. Analysis of quality characteristics

Considering *credibility* and *traceability* the data flow adopted to collect, elaborate and diffuse data is not reported by the analyzed countries with the exception of Italy, where the data flow is partially descripted leaving out information on data collection time periods at local level and their submission to the relevant region. The feature of *currentness* and in particular data *timeliness* represents one of the positive data quality aspects of COVID-19 datasets. Data are mainly daily updated in all countries at local and national level. On the contrary, datasets lack of data *understandability* as all countries report both the name and the values of each variable in their own originated language making it necessary to translate them before integration. Also the absence of the formulas that clearly describe how each indicator is computed makes the comparability of data particularly complex. Moreover, the level of data *disaggregation* is an important feature to be considered as it allows to compare data across countries and to provide a coherent analysis at European level. With the exception of Italy, the other countries analyzed provide data distributed by gender and age ranges.

4. Discussion and Conclusions

The paper presents an analysis of the FAIRness level of datasets distributed by national authorities to map the spread of COVID-19 in six European countries. Moreover, FAIR principles have been conceptually linked with ISO 25012 considering in particular the characteristics of the *inherent data quality.* This was done to explore whether the minimum set of data description identified by the high level, disciplinary-independent FAIR principles cover the main quality features of data. This extended analysis is particularly important considering the crucial role played by the diffusion of COVID-19 analyses on which researchers and policy makers have relied to face pandemic.

Considering FAIR principles, differences across datasets have been detected in their accessibility and findability. The adoption of GitHub services or customized APIs facilitates the access to data and metadata improving their retrievability thanks to standardised, open and universally implementable communications protocols. Moreover, this solution simplifies the assignment of global unique and persistent identifiers to both data and metadata. Conversely, considering the interoperability and reusability principles, all datasets lack the use of a data model as well as of standards for the representation of data and metadata. Moreover, the absence of a clear data flow that describes the provenance of data makes it difficult to integrate data and perform a multi-country analysis. Positively, data are open and may be reused for statistical purposes without requiring authentication to relevant websites.

From a data quality perspective, the attention has been posed on the *inherent data quality* characteristics of ISO 25012. All datasets positively met the feature of currentness with information updated daily at local and national level. This is an important step forward that may be also applied for routinely datasets, as generally medical data are provided one or two years after the collection, making it difficult for scientists to produce innovative and non-obsolete analyses. On the contrary, datasets lack of understandability as no detailed information are reported in terms of indicator definition and formula adopted to compute it. Moreover, the lack of data flow describing its collection, elaboration, aggregation and diffusion makes datasets hard to be accurate and traceable. This is also underlined in previous work [8,9] considering regional based systems where the lack of standardized criteria for data collection might influence the count of cases and tests performed. Finally, the majority of countries provide data distributed by territorial, gender and age ranges level. However, a non-homogeneous distribution is present across both indicators and countries analysed. This data quality feature is critically important for the purpose of the datasets as a coherent distribution may allow a cross-country analysis of the COVID-19 diffusion in Europe.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2015;3:1-9.

[2] Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review. JMIR research protocols. 2021;10(2):e22505.

[3] Smerek MM. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. 2015. Available at: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf. Accessed July 28th, 2021.

[4] WHO. Global surveillance for COVID-19 caused by human infection with COVID-19 virus: Interim guidance. Available at: https://apps.who.int/iris/rest/bitstreams/1272502/retrieve. Accessed 14 July 2020. Accessed July 28th, 2021.

[5] ISO/IEC 25012:2008 - Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model, International Organization for Standardization, Switzerland. 2008.

[6] Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic acids research. 2021;49:D1534-D1540.

[7] Chen, H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. IJERPH. 2014;11(5):5170-5207.

[8] Sartor G, Del Riccio M, Dal Poz I, Bonanni P, Bonaccorsi G. COVID-19 in Italy: Considerations on official data. Int J Infect Dis. 2020;98:188-190.

[9] Alamo T, Reina DG, Mammarella M, Abella A. Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. Electronics. 2020;9(5):827.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210817

Fast Healthcare Interoperability Resources (FHIR) in a FAIR Metadata Registry for COVID-19 Research

Sophie Anne Ines KLOPFENSTEIN[a,b,1], Carina Nina VORISEK[a], Aliaksandra SHUTSKO[c], Moritz LEHNE[a], Julian SASS[a], Matthias LÖBE[d], Carsten Oliver SCHMIDT[e] and Sylvia THUN[a]
[a] *Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Germany*
[b] *Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany*
[c] *ZB MED – Information Centre for Life Sciences, Germany*
[d] *Institute for Medical Informatics (IMISE), University of Leipzig, Germany*
[e] *Institute of Community Medicine, University Medicine Greifswald, Germany*

Abstract. Adopting international standards within health research communities can elevate data FAIRness and widen analysis possibilities. The purpose of this study was to evaluate the mapping feasibility against HL7® Fast Healthcare Interoperability Resources® (FHIR)® of a generic metadata schema (MDS) created for a central search hub gathering COVID-19 health research (studies, questionnaires, documents = MDS resource types). Mapping results were rated by calculating the percentage of FHIR coverage. Among 86 items to map, total mapping coverage was 94%: 50 (58%) of the items were available as standard resources in FHIR and 31 (36%) could be mapped using extensions. Five items (6%) could not be mapped to FHIR. Analyzing each MDS resource type, there was a total mapping coverage of 93% for studies and 95% for questionnaires and documents, with 61% of the MDS items available as standard resources in FHIR for studies, 57% for questionnaires and 52% for documents. Extensions in studies, questionnaires and documents were used in 32%, 38% and 43% of items, respectively. This work shows that FHIR can be used as a standardized format in registries for clinical, epidemiological and public health research. However, further adjustments to the initial MDS are recommended – and two additional items even needed when implementing FHIR. Developing a MDS based on the FHIR standard could be a future approach to reduce data ambiguity and foster interoperability.

Keywords. Metadata Standards, COVID-19, FAIR data, HL7 FHIR, Fast Healthcare Interoperability Resources, Syntactic Interoperability, Infrastructure

1. Introduction

The NFDI4Health Task Force Covid-19 (TF C19) is a project conducted by partners of the National Research Data Infrastructure for Personal Health Data (NFDI4Health). TF C19 aims to develop a FAIR (findable, accessible, interoperable and reusable) data

[1] Corresponding Author, Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité – Universitätsmedizin and Institute of Medical Informatics, Charité – Universitätsmedizin, Charitéplatz 1, 10117 Berlin, Germany; E-mail: sophie.klopfenstein@charite.de

infrastructure for COVID-19 research in Germany and to foster cooperation between clinical, epidemiological and public health communities [1,2]. To gather information from different health data sources on COVID-19 (studies, questionnaires and documents), a metadata schema (MDS) was created and published [3,4].

Fast Healthcare Interoperability Resources ® (FHIR®) is a standard introduced in 2011 by Health Level Seven International (HL7®). It is used in health information technology and provides an information model that is composed of various distinct blocks of information, called resources. Resources intend to provide a definition of the structure and content to cover the information needs of most health information systems. Information not covered by the core resource data model can be captured by an extension mechanism allowing to store and exchange additional structured data. References are used to link resources to each other, while profiles define further rules and constraints on top of standard resources. As FHIR complies with reusability, composability, scalability, performance, usability, data fidelity and implementability principles, it is worthwhile to investigate supporting FHIR in a system [5]. FHIR is mainly used in clinical care, but there are also uses in health research [6,7] and a clinical trials registry [8]. To date there is no FHIR based common registry to gather health data and improve cooperation between clinical, epidemiological and Public Health domains.

Therefore, this paper investigates the feasibility of mapping the MDS to the FHIR standard to enable syntactic and semantic interoperability for NFDI4Health.

2. Methods

Items from the NFDI4Health TF COVID-19 MDS [4] were mapped for each MDS resource type (study, questionnaire, document) to the FHIR standard. The MDS contains two types of items, depending on the resource type they apply to: general and studies specific items. General MDS items were each mapped to the FHIR resources ResearchStudy, Questionnaire and DocumentReference while studies specific items were only mapped to ResearchStudy using FHIR resources of the most current version of HL7 FHIR Version Release 4 (FHIR® R4, v4.0.1) as mapping target [9]. Two mappers (SK, CV), both medical doctors with experience in FHIR, performed the MDS-to-FHIR resource mappings independently after analysis of each MDS item. Incongruities were discussed and solved within a larger mapping team (SK, CV, MLÖ, MLE, ST) resulting in a consolidated version of the mapping, followed up by a feasibility analysis. In some cases, this required further input from a FHIR expert (JS), or MDS expert (AS). Evaluation of the FHIR mapping was done by calculating the percentage of mapping in each category for each MDS resource type and across resource types. Mapping results were categorized based on previous literature as follows [10]: 1) Available as standard resource, 2) Available as extension, 3) Mapping to FHIR not possible

3. Results

Forty-four distinct items from the MDS were split into general (n = 21) and studies specific items (n = 23) resulting in a total of eighty-six items to map [11]. Details on MDS resource types, FHIR resources and availability in FHIR can be found in **Table 1**.

Table 1. NFDI4Health TF COVID-19 MDS-FHIR mapping, N (%)

MDS resource types	**FHIR resources**	**Available as standard resource**	**Available as extension**	**Mapping to FHIR not possible**	**MDS items**
Studies	ResearchStudy	27 (61)	14 (32)	3 (7)	44
Questionnaires	Questionnaire	12 (57)	8 (38)	1 (5)	21
Documents	DocumentReference	11 (52)	9 (43)	1 (5)	21
	Total	50 (58)	31 (36)	5 (6)	86

Among all 86 mapped items, 50 (58%) were available in FHIR as standard resources. Further 31 items (36%) were available as extensions. Five (6%) of the MDS items could not be mapped to FHIR. Analyzing the mapping across MDS resource types, 94% of the MDS items could be mapped either with standard FHIR resources or extensions. Mapping was possible for 93% of the MDS items for studies, and 95% for questionnaires and documents, respectively. **Figure 1** illustrates the availability of MDS items in FHIR.

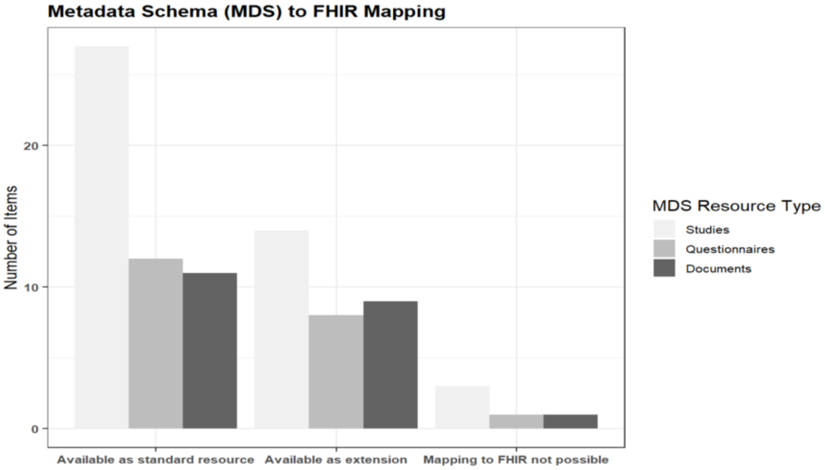


Figure 1. Number of mapped items by MDS resource type and mapping categories.

Some MDS items could not be mapped to FHIR. A mapping of the MDS items "study_status" and "study_analysis_unit" with their respective corresponding mandatory FHIR elements "ResearchStudy.status" and "Group.type" (ResearchStudy is referencing to the Group resource via ResearchStudy.enrollment) was not possible due to incompatible differences in each required value set. Building an extension would not permit to obviate the use of the value sets because of their binding strength. The MDS item "resource_type", relevant for all MDS resources, was not mapped since the selection of the MDS item "resource_type" is followed by a conditional metadata mapping to the appropriate FHIR resource.

In order to use the FHIR resources Questionnaire and DocumentReference, the following items had to be added to the MDS due to the FHIR cardinality: "Questionnaire.status", "DocumentReference.status".

4. Discussion

Existing FHIR resources guarantee the coverage of common requirements but can be expanded in most of the cases using customized extensions [5]. We were able to map the majority of MDS items to FHIR either by using standard resources or custom extensions (94%) demonstrating the flexibility of this standard and its suitability to our use case. However, FHIR resources are designed based on the 80/20 rule (20% of requirements satisfying 80% of the use cases), as well as on reusability and composability principles. With a greater number of extensions needed (36% of the items), we might also lose the proximity to the standard and hinder the compatibility with other systems. Furthermore, analysis of the MDS items showed for example that some definitions are still ambiguous. In some cases, different concepts are covered in a unique item which might lead to complex conditional mappings. Additionally, mapping of two items was not possible because of incompatible value sets between MDS and FHIR. Adjustments of the MDS are recommended and in some cases even needed to ensure compatibility with FHIR. Generating a metadata schema based on FHIR would allow an easier integration of further standards used in various research communities while lowering the amount of FHIR extensions. The current FHIR ResearchStudy resource has a low maturity level (i.e., future changes to this resource are likely) and has a focus on clinical trials. In 2022 HL7 will release a new FHIR version. Previews of the next version show that the ResearchStudy resource will be suitable for studies beyond clinical trials [12]. Therefore, future mappings within our use case could be even more feasible. However, the exact release date is not known and main German health initiatives and projects such as the Medical Informatics Initiative Germany are using FHIR R4 [13] and compatibility is one major aspect within our the NFDI4Health initiative. Further developments should also target bridges to the OMOP data model with its focus on research databases [14].

5. Conclusions

The NFDI4Health TF C19 metadata schema supports a FHIR mapping and therefore can be used for different types of health resources from different research communities. A mapping of the MDS using FHIR standard resources and elements was feasible in more than half of the cases. In most of the cases where FHIR standard elements were not available, FHIR extensions were used. Five items could not be mapped and made MDS adjustments necessary. By creating a COVID-19 registry supporting FHIR, collection of structured data, findability and analysis could be leveraged in different health research communities. We plan to profile FHIR resources based on the mapping of the next metadata schema version (currently under development) and implement the created profiles. For the main project of NFDI4Health, we plan to use FHIR as a basis for a new common metadata schema, enabling syntactic interoperability and facilitating the seamless integration of further standards to ensure semantic interoperability.

Acknowledgments and Competing interests

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche

Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 451265285, PI 345/17-1, SCHM 2744/9-1. | ST is vice chair of HL7 Deutschland e.V.

References

[1] NFDI4Health. What is NFDI4Health? [Internet]. 2020 [cited 2021 Aug 02]. Available from https://www.nfdi4health.de

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

[3] Schmidt CO, Fluck J, Golebiewski M, Grabenhenrich L, Hahn H, Kirsten T, Klammt S, Löbe M, Sax U, Thun S, Pigeot I; NFDI4Health Task Force Covid-19. COVID-19-Forschungsdaten leichter zugänglich machen – Aufbau einer bundesweiten Informationsinfrastruktur [Making COVID-19 research data more accessible-building a nationwide information infrastructure]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2021 Jul 23:1–9. German. doi: 10.1007/s00103-021-03386-x. Epub ahead of print. PMID: 34297162; PMCID: PMC8298983.

[4] Schmidt CO, Darms J, Shutsko A, Löbe M, Nagrani R, Seifert B, Lindstädt B, Golebiewski M, Koleva S, Bender T, Bauer CR, Sax U, Hu X, Lieser M, Junker V, Klopfenstein S, Zeleke A, Waltemath D, Pigeot I, Fluck J; NFDI4Health Task Force COVID-19. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Stud Health Technol Inform. 2021 May 27;281:794-798. doi: 10.3233/SHTI210284. PMID: 34042687.

[5] HL7. 2.16 FHIR Overview – Architects. [Internet]. 2019 [cited 2021 Aug 02]. Available from https://www.hl7.org/fhir/overview-arch.html

[6] Gruendner J, Wolf N, Tögel L, Haller F, Prokosch HU, Christoph J. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): System Design and Implementation. J Med Internet Res. 2020 Oct 7;22(10):e19879. doi: 10.2196/19879. PMID: 33026356; PMCID: PMC7578821.

[7] Lee HA, Kung HH, Lee YJ, Chao JC, Udayasankaran JG, Fan HC, Ng KK, Chang YK, Kijsanayotin B, Marcelo AB, Hsu CY. Global Infectious Disease Surveillance and Case Tracking System for COVID-19: Development Study. JMIR Med Inform. 2020 Dec 22;8(12):e20567. doi: 10.2196/20567. PMID: 33320826; PMCID: PMC7758088.

[8] Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping. Stud Health Technol Inform. 2018;253:3-7. PMID: 30147028.

[9] HL7.org. HL7® FHIR® Release 4. [Internet]. 2019. [cited 2021 Aug 02]. Available from http://hl7.org/fhir/R4/index.html

[10] Garza MY, Rutherford M, Myneni S, Fenton S, Walden A, Topaloglu U, Eisenstein E, Kumar KR, Zimmerman KO, Rocca M, Gordon GS, Hume S, Wang Z, Zozus M. Evaluating the Coverage of the HL7 ® FHIR ® Standard to Support eSource Data Exchange Implementations for use in Multi-Site Clinical Research Studies. AMIA Annu Symp Proc. 2021 Jan 25;2020:472-481. PMID: 33936420; PMCID: PMC8075534.

[11] NFDI4Health Task Force COVID-19 Metadata Schema (MDS) Mapping to FHIR - MDS-to-FHIR Mapping [Data Set] [Internet]. FAIRDOMHub; 2021. Available from: https://fairdomhub.org/data_files/4210?version=2

[12] HL7. HL7® FHIR® Release 5 Draft Ballot – 8.22 Resource ResearchStudy. [Internet]. 2021. [cited 2021 Aug 04]. Available from https://build.fhir.org/researchstudy.html

[13] SIMPLIFIER.NET. Medizininformatik Initiative. [Internet]. 2021. [cited 2021 Sep 16.] Available from https://simplifier.net/organization/koordinationsstellemii/~projects

[14] Rinaldi E, Thun S. From OpenEHR to FHIR and OMOP Data Model for Microbiology Findings. Stud Health Technol Inform. 2021 May 27;281:402-406. doi: 10.3233/SHTI210189. PMID: 34042774.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210818

Improving the FAIRness of Health Studies in Germany: The German Central Health Study Hub COVID-19

Johannes DARMS[a,1], Jörg HENKE[b], Xioaming HU[c], Carsten Oliver SCHMIDT[b], Martin GOLEBIEWSKI[c] and Juliane FLUCK[a] on behalf of the NFDI4Health Task Force COVID-19
[a] *ZB MED – Information center for Live Sciences, Germany*
[b] *University Medicine of Greifswald, Germany*
[c] *Heidelberg Institute for Theoretical Studies, Germany*

Abstract. The German Central Health Study Hub COVID-19 is an online service that offers bundled access to COVID-19 related studies conducted in Germany. It combines metadata and other information of epidemiologic, public health and clinical studies into a single data repository for FAIR data access. In addition to study characteristics the system also allows easy access to study documents, as well as instruments for data collection. Study metadata and survey instruments are decomposed into individual data items and semantically enriched to ease the findability. Data from existing clinical trial registries (DRKS, clinicaltrails.gov and WHO ICTRP) are merged with epidemiological and public health studies manually collected and entered. More than 850 studies are listed as of September 2021.

Keywords. COVID-19, FAIR, study data portal

1. Introduction

A quickly growing number of clinical trials, as well as public health and epidemiological studies on COVID-19 have started and are already ongoing, but there is a lack of coordination among these efforts for securing common standards, comparable results, and – most importantly – unified access to these results.

Registries such as the German Clinical Trials Register (DRKS) and clinicaltrail.gov collect and provide information about planned, running, and completed studies. Thereby registries help researchers to find studies they are interested in. However, some studies are registered in multiple portals, while others are not registered at all. That is especially true for observational studies without any formal obligation to be registered. To our knowledge, there is no portal that provides an overarching search for clinical trials as well as for epidemiological and public health studies. In addition, current registries are commonly limited to describe overall study characteristics; detailed information about the collected data elements is lacking as well as additional study documents. This may

[1] Corresponding Author, Johannes Darms, ZB MED – Information Centre for Life Sciences, Gleueler Straße 60, 50931 Cologne, Germany, Germany; E-mail: darms@zbmed.de.

impair the possibility to find the desired information of interest. Services are needed to further facilitate the findability of studies.

The NFDI4Health Task Force COVID-19 initiative [1] was established to address those issues and increase the FAIRness – Findability, Accessibility, Interoperability and Reusability – of clinical, epidemiologic and public health studies with a COVID-19 focus. Therefore, the consortium has developed the German Central Health Study Hub Covid-19, a webservice to search for COVID-19 related studies in Germany. This service provides an overview of existing clinical trial registries (DRKS, clinicaltrails.gov and WHO ICTRP) and also includes epidemiological and public health studies that have been manually collected and entered.

2. Methods

We have combined existing platforms from different domains within the German Central Health Study Hub Covid-19. The SEEK platform [2], developed by the FAIRDOM initiative, is mainly used to store study-level metadata, as well as documents and other resources of the studies with their metadata and makes them accessible. In addition, the SEEK system is used to register data and documents with Digital Object Identifiers (DOIs). In contrast, the software systems OPAL/MICA [3] provide search and comparison techniques for data items mainly of survey instruments. OPAL provides access to characteristics of study instruments (such as labels, value lists, missing definitions and annotations). MICA allows browsing variable definitions and related studies. The variable search is enriched by semantic annotations of the Maelstrom Taxonomy [4].

To provide a unified user interface with a dedicated look & feel a new webpage has been developed. The system is a single-page application developed as a React application that combines selected information stored in SEEK, MICA/OPAL into a simplified search interface. To increase the findability and accuracy of search queries, semantically enriched information is stored in a dedicated search instance (elasticsearch).

A data model was developed to capture and harmonize information from the different sources and improve findability. The model is based on attributes used by clinicaltrails.gov [5], DRKS [6], and WHO ICTRP [7]. In addition, the required properties for assigning DOIs have been added by adhering to the DataCite scheme [8]. Those properties are needed to capture metadata about associated documents such as questionnaires, data dictionaries, and eCRFs.

A major difference of the developed data model from many others is the ability to describe a hierarchy between studies and associated documents. For example, one can model that a survey instrument is used by multiple studies (indicating reuse of data collection forms) or that one study is part of another study. A more detailed description of the data model including the set of minimal required properties and software components as well their interconnection can be found in [9]. The minimum dataset that must be included within the platform is influenced by the DataCite Schema, as some properties are required for DOIs to be assigned. In addition, some properties (title, description and study status and primary design) are mandatory for studies. The entry of other relevant metadata is recommended but not mandatory to keep the entry barrier for authors low.

The software system also includes a procedure for de-duplication of information. When duplicate resources are detected, the version from the data source with the highest priority is selected. The order is chosen by similarity with our data model i.e., more fields are equivalent. Duplicates can occur since some studies are registered in multiple registers. Especially within the WHO ICTRP dataset as it aggregates studies form other registers. The following priority list is used to resolve the problem: manually collected information, clinicaltrails.gov, DRKS, WHO ICTRP.

While not part of the software, the process to integrate study descriptions and associated documents is equality important. A business process to collect and integrate information has been designed (publication in preparation). The process is largely performed by trained data stewards. The process starts with a search for public information about a study. If some information can be obtained, it is collected and a template pre-filled with this information is sent to the study authors, otherwise a blank template is sent. Supportive mail/phone exchanges are used to assist the study author in providing needed information. When the related study documents are made available, assistance is provided in selecting the correct license.

3. Results

Our COVID-19 study hub improves the findability and accessibility of clinical, epidemiologic and public health studies related to this topic and, thus enhances their FAIRness as some of the 25 manually collected observational studies were previously not listed in any portal. The initial focus is on studies in Germany and international studies with German contributions. However, this infrastructure can also be helpful for bundling information, metadata and resources of studies in other countries or internationally, as the underlying metadata structures are generic.

Content was obtained in two ways: either by reusing existing information or by querying information directly from studies of interest, that have been identified based on a predefined requirements catalogue. Integration from existing registers (DRKS, clinicaltrails.gov and WHO ICTRP) is done automatically, but a conversion between data formats is needed. Some values may not be transferred, and others may not be provided. The final step is automatic deduplication by removing studies that are listed in multiple registers. On the other hand, the manual process of asking studies for information is labor intensive but may result in more comprehensive information in alignment with our requirements on attributes related to the study.

The study documents and resources stored in the SEEK component include study-protocol templates and data dictionaries as well as information on study-metadata structures – such as data models that describe study subjects and their clinical parameters – in addition to treatment outcomes and similar information. Additionally, direct links to primary resources and websites for the studies are included. These study information, resources and metadata can be directly searched, browsed and accessed. The MICA component of the system helps the users to find specific variables within survey instruments of interest. MICA allows to select and filter variables from the available studies to compare variable definition and its attributes.

As of September 2021, the system contained information from over 850 COVID-19 studies (46 manually collected, 158 obtained through WHO ICTRP, 468 through DRKS and 202 through NCIT) most of which are conducted in Germany. Some of the staff responsible for studies shared documents of relevance such as data collection tools, i.e., data dictionaries, questionnaires, and eCRFs. 23 data collection instruments are described at the level of individual data elements (i.e., questions, data properties), including a semantic annotation to better compare covered areas within and across instruments. The system does not contain privacy sensitive information.

The German Central Health Study Hub COVID-19 is freely accessible under https://covid19.studyhub.nfdi4health.de. The platform has already been accessed by more than 200 unique visitors a month and receives around 500 requests per day. All content can be accessed via web-interfaces and some parts are also accessible via web services (API). The software system, based on the 3 interlinked components SEEK, MICA and frontend search interface, enables browsing, accessing and comparison of COVID-19 studies and their descriptive metadata, their data collection elements, as provides search functions for studies, data collection instruments and elements, as well as related documents.

4. Discussion

We released a service to increase the FAIRness of COVID-19 related studies in Germany. The service reuses and combines existing technologies and widely used data management platforms with a sophisticated metadata schema. Data are collected and entered manually from studies (especially for epidemiological and public health studies), as well as automatically captured and reused where possible e.g., for data from clinical trials. Many interventional as well as non-interventional studies have already been published in registries such as clinicaltrails.gov and DRKS. Many studies listed in our system are taken from there. We considered the aggregation of this information as a benefit of our service. The integration of data collection instruments and item banks adds to the functionality. Item deconstruction of survey instruments and their semantic enrichment is also available in the MDM portal [10]. However, to our knowledge, there is no service that provides unified access to studies and their decomposed survey instruments.

Semantic enrichment was performed with the Maelstrom Taxonomy to ease the search for relevant information. Recent works conducted in the consortia [11] showed that SNOMED is also suited as a basis to semantically enrich data collection instruments. Therefore, we currently elaborate to do so to further increase the reusability and interoperability of the collected data.

The data schema was developed to meet the various requirements. As we could not directly use an existing schema and had to create a new one. One rationale to proceed as we did is our emphasis on fast development, due the rapid spread of the disease. However, we ensured compatibility with the current FHIR specification [12]. Our next task is to create FHIR profiles/extensions to convert our schema into a standardized format to increase the interoperability of the contained data.

The reuse of documents such as survey instruments is often hindered by legal restrictions. This problem occurs when collecting and publishing data collection instruments, there copyrights can and are claimed. Therefore, prior to inclusion in our FAIR platform, copyrights must be clarified, and documents must be licensed under some appropriate open license, such as a Creative Commons license. However, this process is not straightforward and delays the integration of instruments into the portal.

5. Conclusion

We have established a service to increase the FAIRness of clinical, epidemiologic and public health studies and associated documents. The harmonization of existing information and integration of previous unavailable information, as well as semantic enrichment of information eases the findability of COVID-19 related studies conducted in Germany. In order to further increase usefulness of the service i.e., the number of studies included, procedures to simplify the process of (meta) data collection are in preparation. The first is an interactive web-based user-form to will facilitate study registration. Furthermore, the business process used by data stewards to collect information is currently being streamlined and will be supported by the software stack. Additionally, usability is being evaluated to guide further development of the software system to meet user needs.

Acknowledgements

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de/task-force-covid-19-2). We gratefully acknowledge the financial support of the German Research Foundation (DFG) – Project Number 451265285, and from the Klaus Tschira Foundation (KTS).

References

[1] Task Force COVID-19 Team. Task Force COVID-19 - NFDI4Health [Internet]. [cited 2021 Aug 6]. Available from: https://www.nfdi4health.de/de/task-force-covid-19

[2] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, et al. SEEK: a systems biology data and model management platform. BMC systems biology. 2015;9(1):1–12.

[3] Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. International journal of epidemiology. 2017;

[4] Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS One. 2018;13(7):e0200926.

[5] National Library of Medicine, National Institutes of Health. XML Schema for ClinicalTrials.gov public XML [Internet]. [cited 2021 Aug 6]. Available from: https://clinicaltrials.gov/ct2/html/images/info/public.xsd

[6] DRKS. Description of entry fields [Internet]. [cited 2021 Aug 6]. Available from: https://www.drks.de/drks_web/navigate.do?navigationId=entryfields&messageDE=Beschreibung%20der%20Eingabefelder&messageEN=Description%20of%20entry%20fields

[7] WHO. World Health Organisation - ICTRP Search Portal [Internet]. [cited 2021 Aug 6]. Available from: https://www.who.int/clinical-trials-registry-platform/the-ictrp-search-portal

[8] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data. Version 4.3 [Internet]. [cited 2021 Aug 6]. Available from: https://schema.datacite.org/meta/kernel-4.3/

[9] Schmidt CO, Darms J, Shutsko A, Löbe M, Nagrani R, Seifert B, et al. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Studies in Health Technology and Informatics. 2021;281:794–8.

[10] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database. 2016;2016.

[11] Vorisek CN, et al. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies. In: Public Health and Informatics. IOS Press; 2021. p. 88–92.

[12] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. IEEE; 2013. p. 326–31.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210819

From EHR to EDC - The Experience at the Policlinico Hospital in Milan

Sara PIZZIMENTI [a,1] , Mauro BUCALO [b], Amedeo GUZZARDELLA [a], Eleonora FERRETTI [a], Angelo CAROLI [a], Alberto ZANELLA [a], Giacomo GRASSELLI [a], Nicola BARBARINI [b] and Silvano BOSARI [a]
[a] *Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy*
[b] *BIOMERIS (BIOMEdical Research Informatics Solutions), Pavia, Italy*

Keywords. data reuse, secondary use, i2b2, REDCap, eCRF, EDC

1. Introduction

As in many hospitals, especially in Italy, the Electronic Health Record (EHR) of IRCCS Ca' Granda Ospedale Maggiore Policlinico hospital in Milan is composed by a variety of proprietary software applications used to support clinical practice. For this reason, clinical data reuse or secondary use for research purposes is a difficult goal to achieve. Our hospital faced this problem by i2b2 (Informatics for Integrating Biology and the Bedside) [1], a data warehouse that aggregates EHR heterogenous data following the FAIR principles, making them easy to access and find by researchers but also reusable and interoperable with other systems. i2b2 system includes currently 4 million patients and 520 million observations, now available for research purposes, such as automatically filling an EDC (Electronic Data Capture) system for clinical studies [2]. The EDC software in use within our hospital is REDCap (Research Electronic Data Capture) [3], which is accessible through Application Programming Interfaces (APIs) that allow data import from external sources. The aim of this work is to evaluate a procedure for the automatic import of data from clinical practice to an EDC system for a clinical study.

2. Methods

i2b2 system within the Policlinico Hospital in Milan aggregates heterogenous data sources and the i2b2 table Patient Mapping contains the lookup between the patient identifier of each data source and the unique i2b2 patient id.

We developed an ETL (Extract, Transform, Load) procedure for importing data from i2b2 system to a REDCap study. The ETL procedure is based on JavaScript code, which runs on Mirth Server, from which HTTP POST requests containing data to be imported are sent using the REDCap API. Preliminary operation of the procedure includes the mapping of the REDCap variables into the i2b2 concepts. Then the mapping table between i2b2 and REDCap is created by joining i2b2 patient mapping table with

[1] Corresponding Author, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Via Francesco Sforza, 35 – 20122 Milan, Italy; E-mail: sara.pizzimenti@policlinico.mi.it.

the enrollment list containing the patient REDCap project identifier and the identity information. i2b2 queries are implemented for each data collection event in the REDCap project to extract data from i2b2 and then import into REDCap. The main query parameters are the patient identifier, the event date and the concept related to REDCap variables. Dates can be obtained directly from the value of a variable manually entered or automatically calculated starting from one or more reference dates within the project.

3. Results

The impact of automated import procedures has been evaluated on a multicenter study promoted by Policlinico Hospital and approved by Ethical Committee. The study aims to collect data from patient admitted to intensive care unit with a COVID19 infection. Our hospital has currently collected data about 279 patients. This study started in 2020 following the health emergency and patient data has been collected by hand ever since.

We decide to evaluate the import procedures by reusing structured data already stored in the hospital. In particular, we considered 31 laboratory tests collected in different events: on the first day of admission and the following hospitalization days, a total of 2482 REDCap events for the whole cohort. We compared the values entered manually with the automatically imported ones for a total number of 76942 (31 laboratory test values for each REDCap event). The comparison with a 5% of value tolerance shows that 93,5% of the manually entered exams in REDCap were found also by the automatic procedure; the remaining values could be due to manual entry errors on the exam's values or on the associated dates. Furthermore, 62,8% of the 36018 missing values were filled in by the automatic procedure. Missing values have different reasons including the onerous request for time for manual data entry. Finally, considering the total of 63530 of exams found by almost one of the two approaches, 62% come from both systems, while the 36% was entered only by automatic procedure.

4. Conclusions

Automatic laboratory tests results import led to improvements in terms of time, accuracy, and quality, showing that the automatic procedure should have applied from the beginning of data collection. For these reasons we plan to apply such a procedure on other studies and expanding the coverage of imported data types.

References

[1] Kohane IS Churchill SE Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 2012; 19:181–5.

[2] Griffon N, Pereira H, Djadi-Prat J, García MT, Testoni S, Cariou M, Hilbey J, N'Dja A, Navarro G, Gentili N, Nanni O, Raineri M, Chatellier G, Gómez De La Camara A, Lewi M, Sundgren M, Daniel C, Garvey A, Todorovic M, Ammour N. Performances of a Solution to Semi-Automatically Fill eCRF with Data from the Electronic Health Record: Protocol for a Prospective Individual Participant Data Meta-Analysis. Stud Health Technol Inform. 2020 Jun 16;270:367-371.

[3] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)-A Metadata-driven Methodology and Workflow Process for Providing Translational Research Informatics Support. Journal of Biomedical Informatics. 2009 Apr; 42(2):377–381.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210820

Towards FAIR Patient Reported Outcome: Application of the Interoperability Principle for Mobile Pandemic Apps

Michael Rusongoza MUZOORA[a,b,1], Marco SCHAARSCHMIDT[a,b],
Dagmar KREFTING[c,d], Johannes OEHM[e], Sarah RIEPENHAUSEN[e] and
Sylvia THUN[b,f]
[a] *Berlin Institute of Health (BIH), Germany*
[b] *Charite` – Universita¨tsmedizin Berlin, Germany*
[c] *Department of Medical Informatics, University Medical Center Gottingen (UMG), Germany*
[d] *Campus-Institute of Data Science (CIDAS) Gottingen, Germany*
[e] *Institute of Medical Informatics, University of Munster, Germany*
[f] *Hochschule Niederrhein - University of Applied Sciences, Krefeld, Germany*

Keywords. Pandemic App, COVID-19, Interoperability, Healthcare Data

1. Introduction

During the COVID-19 pandemic several individual apps have been developed to collect and track medical data. Although many of them address similar or at least overlapping aspects, due to diverse data models and formats, data items cannot be jointly analysed to increase evidence [1]. The NUM-COMPASS project [2], which is part of the German COVID-19 Research Network of University Medicine (NUM), built a coordination and technology platform as starting point for researchers and app developers. It enables them to collect data compliant to the German Corona Consensus Dataset (GECCO) [3]. This addresses in particular interoperability as part of the FAIR guiding principles [4]. This paper describes the implementation of these principles by NUM-COMPASS, to support joint analyses on shared data collected by various app-based studies.

2. Methods and Results

Interoperability is of particular importance, as research data often comes from multiple sources. It needs to be integrated into existing systems for analysis or processing. The following aspects have been defined for interoperability within the FAIR principles [4]:

- I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

[1] Corresponding Author, Michael Rusongoza Muzoora, Berlin Institute of Health, Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany; E-mail: michael.muzoora@charite.de.

- I2: (Meta)data use vocabularies that follow FAIR principles
- I3: (Meta)data include qualified references to other (meta)data

In the following, the measures to assure interoperability between pandemic (and further) apps are described. We interpret **I1** as compliance to syntactic interoperability, while **I2** relates to semantic interoperability. **I3** is in particular relevant for the contextual information about the data.

I1: COMPASS apps use HL7 FHIR as language and JSON as interchange format. The GECCO data model itself [3] is defined by the FHIR profiles and the FHIR resources used. Thus, GECCO merges into existing data models, which is also the case for the collected data in COMPASS. These all provide a well-defined (meta)data model structure, which can be validated by an automated conformity check.

I2: To adhere to I2, internationally recognized controlled terminologies, ontologies and thesauri have been used in the underlying FHIR profiles, such as LOINC, SNOMED and ICD-10 GM.

I3: FHIR inherently fulfills this requirement, as links to other FHIR resources and vocabularies are a basic concept in FHIR. The GECCO data model comprises of several interlinked FHIR resources. In Germany, the Medical Informatics Initiative (MII) [5] provides the reference links for the vocabularies as well as for FHIR profiles of the nation- ally consented core data set. For billing in the German health care system, the coding of the ICD-10-GM is mandatory as per Law §§ 301, 295 Sozialgesetzbuch SGB V [6]. However, it is not sufficient for some cases. Therefore, the use of further terminologies (SNOMED CT, ORPHA codes) is encouraged by the MII.

3. Conclusion

The GECCO data model has been successfully integrated into the COMPASS app frame- work to improve interoperability. An automated conformity check is freely accessible to help researchers to assess if the data can be integrated into a common analysis from the beginning on. As a result, a compliance seal is provided. However, with fluctuant research questions, a general process to provide interoperable data beyond GECCO is currently developed.

References

[1] Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med. 2019;2:79.

[2] Netzwerk Medizin Universita¨ten. (2021, June). Best practices and common solutions for mobile pandemic applications. COMPASS. https://num-compass.science/de/deliverables/

[3] Sass, J., Bartschke, A., Lehne, M. et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 20, 341 (2020). https://doi.org/10.1186/s12911-020-01374-w.

[4] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.

[5] Medizin Informatik Initiative. (2021). Arbeitsgruppe Interoperabilita¨t — Medizininformatik-Initiative. https://www.medizininformatik-initiative.de/de/zusammenarbeit/arbeitsgruppe-interoperabilitaet

[6] Bundesregierung Deutschland. (2021, May 28). §§ 301, 295 SGB V Krankenha¨user. Sozialgesetzbuch (SGB V) Fu¨nftes Buch Gesetzliche Krankenversicherung. https://www.sozialgesetzbuch-sgb.de/sgbv/301.html

Section IV

Metadata, Ontologies and Terminologies to Support Sharing of Health Research Data

This page intentionally left blank

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210822

Automated Modeling of Clinical Narrative with High Definition Natural Language Processing Using Solor and Analysis Normal Form

Melissa P. RESNICK[a,1], Frank LeHOUILLIER[a], Steven H. BROWN[b], Keith E. CAMPBELL[b], Diane MONTELLA[b] and Peter L. ELKIN[a,b,c,d]
[a] *Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA*
[b] *U.S. Department of Veterans Affairs, Office of Health Informatics, USA*
[c] *U.S. Department of Veterans Affairs, WNY VA, USA*
[d] *Faculty of Engineering, University of Southern Denmark, Denmark*

Abstract. Objective: One important concept in informatics is data which meets the principles of Findability, Accessibility, Interoperability and Reusability (FAIR). Standards, such as terminologies (findability), assist with important tasks like interoperability, Natural Language Processing (NLP) (accessibility) and decision support (reusability). One terminology, Solor, integrates SNOMED CT, LOINC and RxNorm. We describe Solor, HL7 Analysis Normal Form (ANF), and their use with the high definition natural language processing (HD-NLP) program. Methods: We used HD-NLP to process 694 clinical narratives prior modeled by human experts into Solor and ANF. We compared HD-NLP output to the expert gold standard for 20% of the sample. Each clinical statement was judged "correct" if HD-NLP output matched ANF structure and Solor concepts, or "incorrect" if any ANF structure or Solor concepts were missing or incorrect. Judgements were summed to give totals for "correct" and "incorrect". Results: 113 (80.7%) correct, 26 (18.6%) incorrect, and 1 error. Inter-rater reliability was 97.5% with Cohen's kappa of 0.948. Conclusion: The HD-NLP software provides useable complex standards-based representations for important clinical statements designed to drive CDS.

Keywords. Natural Language Processing, Interoperability, Clinical Decision Support, Controlled Terminology

1. Introduction

Technical (syntactic) interoperability addresses how computers exchange data. This is accomplished with messaging protocols and data formats. Semantic interoperability builds upon syntactic interoperability and addresses how computers interpret meaning of data. Semantic interoperability allows EHRs to unambiguously and consistently determine meaning of the data for data presentation and decision support.

[1] Corresponding Author, Melissa P. Resnick, Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, 14203, United States of America; E-mail: mresnick@buffalo.edu.

Terminologies, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), are used as data encoding standards in informatics and contribute to semantic interoperability. These terminologies can also provide a foundation for other tasks such as knowledge management, data integration, and decision support[1]. Three of the most commonly used terminology standards are SNOMED CT, LOINC and RxNorm, each with particular strengths. Increasingly, there is interest in combining these partially overlapping standards to enhance clinical expressivity. Solor is an integrated terminology system created in collaboration with the U.S. Veterans Affairs (VA) [1] that combines SNOMED CT (representing diseases, findings, and procedures), Logical Observation Identifiers, Names, and Codes (representing laboratory test results), and RxNorm (representing medications) [2,3].

1.1. Solor

The Solor terminology layer builds primarily upon SNOMED CT, RxNorm, and LOINC by integrating their content and semantics, and normalizing the means to identify and version components, lexically search, logically define, semantically retrieve, and collaboratively extend. The potential advantages of a computable approach enabled by combining SNOMED CT, LOINC, and RxNorm into a single consistent suite for encoding clinical knowledge and data are clear; clinical data can flow among clinical documentation, decision support applications, and order entry at the point of care. This single consistent method of encoding clinical data can also support research, quality measurement, and other secondary uses.

Solor has two fundamental building blocks: concepts and semantics [4]. A concept is defined as an idea or a medically related idea, such as heart attack [4]. These ideas also include a synonym or a fully specified name. A semantic is data that provides contextual meaning to the concept [4].

Like SNOMED CT, Solor is built on a logic model[3]. Most of the terms are shared by Solor and SNOMED CT and these concepts are arranged into hierarchies using "is a" relationships[4]. Each concept has at least one "is a" relationship, except for the top level concepts, which are the most general concepts. Thus, due to the "is a" relationships, one can traverse the hierarchies from the general concepts to more specific concepts.

One goal of using Solor is to improve interoperability. Interoperability of EHR data is critical for clinical decision support. Given that health care for an individual is often delivered by more than one health provider, integration of data from multiple health providers is needed to view the complete health record. To achieve interoperability, clinical systems must understand both the structure (syntax) and meaning (semantics) of the clinical information being exchanged3. Without these two features, information may be viewable by humans, but not integrated for clinical decision support3. One way for providing interoperability is through the use of standards, such as Solor. Solor allows for interoperability by providing structure and meaning to the patient data being exchanged between health providers.

1.2. ANF

Analysis Normal Form (ANF) is a type of highly regularized small information model that is designed to be independent of the content of the clinical statement. For example, a single ANF "performance statement" model can be used to describe any action that has previously been performed, and – if applicable - the results of that action. Broad classes of actions are represented identically including observations of presence or absence of a clinical phenomenon, undergoing a procedure, or the administration of a medication.

The goal of ANF is to provide a simple, consistent and highly re-usable information model for clinical statements. This makes it easier for analysts to understand the data and how it is stored than requiring knowledge of hundreds or thousands of statement-specific specializations. It also helps to ensure that the data can be expressed in an operable and scalable way. The more that data is normalized, the simpler it will become to analyze, and the likelihood of analysis errors will be reduced. ANF represents clinical data for data analyst's purposes, not in a way we may choose to display the data for a clinician[4]. ANF was approved as an HL7 informational ballot in 2019 [5].

1.3. HD-NLP

High Definition Natural Language Processing (HD-NLP) is a pipeline developed at the University at Buffalo, which evolved from the HTP-NLP work from UB. The system uses a full semantic parse in memory and then uses an encoder to link text to any set of Ontologies which a user wants to use to represent the knowledge in the free text being codified. Each entity is tagged as an affirmed, negative or uncertain assertion and each is further tagged with a date time stamp. We then automatically generate compositional expressions where applicable in the source text using the semantic relations available in the ontologies being encoded. We add the metadata from the record using the analysis normal form standard and link it to the information stored from computing over the input string.

HD-NLP uses several sources of synonymy, kept in separate synonym sets (synsets) which are available for interrogation to understand why certain results were obtained. The system is architected so that the input queries come to an input queue and then they are processed and sent to an output queue where each job can then be picked up by the user. This is available as a web service. The service can provide Solor and ANF output but also can limit its search to the source ontologies (SNOMED CT, RxNorm, and LOINC). We made use of the HD-NLP to rapidly assign terminology concepts to text in patient records or KNART (knowledge artifact) input text [6-8]. A level of syntactic processing was required to match text with ontological terms. The linguistic representation is specified in language models. Of primary concern here was an English language model to identify sentences, phrases, words, and parts of speech. Terms from Solor and its source ontologies were then assigned to spans of text.

2. Methods

Narrative clinical statements designed for clinical decision support numbering 694 were obtained from the VA KNART project to create clinical content using the HL7 Knowledge Artifact specification [9]. Each of these clinical statements were previously assigned Solor concepts in ANF structures by experienced human modelers.

We used HD-NLP to algorithmically assign terminology concepts to text in KNART input text. HD-NLP program output consisted of the input narrative clinical statements and the corresponding ANF/Solor models.

Authors MR and PE reviewed 140 (20%) randomly selected narrative clinical statements with their corresponding HD-NLP outputs and compared them to the human modeled “gold standard”. Each narrative clinical statement was judged as "correct" if the HD-NLP output matched the human modeled Solor concepts and human modeled ANF structure. The output was judged as "incorrect" if either the Solor concepts or ANF structure were missing or incorrect. These were then summed to give a total for "correct" and "incorrect" respectively. Forty of 140 elements were double reviewed and conflicts were resolved by consensus. A kappa interrater reliability statistic was calculated.

3. Results

Of the 140 HD-NLP outputs containing both Solor concepts and ANF structures we found: 26 (18.6%) incorrect outputs, 113 (80.7%) correct outputs, and 1 error. The error was due to the fact that there was no output for that single record, for some unknown reason. Incorrect was triggered mainly by missing Solor concepts in the HD-NLP output. In some cases modifier concepts, such as "alcohol" in "alcohol abuse," "former" in "former illicit substance use," and "cup-to-disc" in "cup-to-disc ratio" were missed. In a couple of cases the output was completely incorrect. For example, the input read "polyp cytology shows high-grade dysplasia," while the output read "polyp aplasia." In addition, difficulties were seen with laterality. For instance, the concept "left" in the input was represented as "left to right" in the HD-NLP output. Despite these difficulties, the HD-NLP output was correct in most cases. These include such examples as follows: (1) input as "regular menstrual cycle" and output as "regular periods," (2) input as "patient gender is female" and output as "patient sex female girl," and (3) input as "cognitive impairment" and output as "cognitive impairment." Inter-rater reliability was 97.5% with a Cohen’s Kappa of 0.948.

4. Discussion

We believe further improvements are possible and needed. This includes improvement in such items as: (1) missing or bad synonymy, and (2) bad laterality mappings.

Solor, a formal integration of SNOMED CT, LOINC and RxNorm, represents a significant advance towards semantic interoperability and health information exchange. In addition, it will improve the findability of important clinical statements. The Analysis Normal Form brings standardization to the small information models needed to complete a clinical statement, while enhancing consistency and reducing complexity.

Natural language processing with HD-NLP can provide access to data by mapping clinical utterances in notes and reports to clinical statements, which are reused for clinical decision support. By modeling the KNARTS with Solor and ANF using the HD-NLP system, we can provide a representation that will match the HD-NLP derived data from clinical notes and reports that can then be used to trigger clinical decision support rules. In addition, we expect that HD-NLP can reduce coding burdens on clinicians during data entry, providing well-coded structured data for CDS. This partnership between standards and technology can assist our ability to make practical clinical decision support which may otherwise require duplicate and structured data entry. The more seamless our CDS implementations are, the more they will be easily implemented and shared, fulfilling the important FAIR principle in informatics.

5. Conclusions

Solor integrates SNOMED CT, LOINC, and RxNorm, not merely by just combining these terminologies, but by using an underlying logic model, improving semantic interoperability. This provides improved findability and reusability of data for clinical decision support. ANF further improves interoperability by providing simple and consistent structure to deliver terminological payload as statement models about patients. The HD-NLP software provides access to important clinical statements, which are required to drive CDS. By using this pipeline in a FAIR manner, we can improve the safety and efficacy of the healthcare that we provide for our patients.

Acknowledgments

This work has been supported in part by grants from NIH NLM T15LM012595, NIAAA R21AA026954, NIAAA R33AA026954, and NCATS UL1TR001412. This study was funded in part by the U.S. Department of Veterans Affairs.

References

[1] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79.

[2] Resnick MP, Brown SH, Campbell KE, Montella D, LeHouillier F, Elkin PL. Turning Data into Information: Evaluation of SOLOR. Poster presented at the: AMIA annual symposium; November 2020. Accessed March 9, 2021. https://knowledge.amia.org/72332-amia-1.4602255/t005-1.4604904/t005-1.4604905/3416966-1.4605215/3416966-1.4605216?qr=1

[3] Staes C, Campbell K. From Retrospective Mapping to Prospective Standardization: A Comparison of Integration Strategies to Achieve Semantic Data Interoperability. Department of Veterans Affairs,Veterans Health Administration (VHA) Office of Informatics and Analytics (OIA) Knowledge Based Systems (KBS); 2017:26. Accessed March 8, 2021. http://solor.io/wp-content/uploads/2017/12/White-paper_Achieving-semantic-data-interoperability.pdf

[4] Sujansky W. ISAAC's KOMET and Solor - A Treatise on Symbolic Data Systems.; 2019:194. http://solor.io/wp-content/uploads/2019/02/symbolic-information-analytics-20190226.pdf

[5] Singnureanu I. HL7 Informative Ballot HL7 CIMI Logical Model: Analysis Normal Form (ANF), Release 1.; 2019:130. http://solor.io/wp-content/uploads/2019/08/ANF_Ballot_20190819.pdf

[6] Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Intern Med. 2012;156(1 Pt 1):11-18. doi:10.7326/0003-4819-156-1-201201030-00003

[7] Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011;306(8):848-855. doi:10.1001/jama.2011.1204

[8] Schlegel DR, Crowner C, Lehoullier F, Elkin PL. HTP-NLP: A New NLP System for High Throughput Phenotyping. Stud Health Technol Inform. 2017;235:276-280.

[9] HL7 Standards Product Brief - HL7 Standard: Clinical Decision Support Knowledge Artifact Specification, Release 1.3 | HL7 International. Accessed March 9, 2021. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=337.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210823

Distribution-Based Similarity Measures Applied to Laboratory Results Matching

Martin COURTOIS[a,1], Alexandre FILIOT[a] and Gregoire FICHEUR[a,b]
[a] CHU Lille, INCLUDE: Integration Center of the Lille University hospital for Data Exploration, F-59000, Lille, France
[b] Univ. Lille, CHU Lille, ULR 2694 - METRICS, Public health dept, F-59000, Lille, France

Abstract. The use of international laboratory terminologies inside hospital information systems is required to conduct data reuse analyses through inter-hospital databases. While most terminology matching techniques performing semantic interoperability are language-based, another strategy is to use distribution matching that performs terms matching based on the statistical similarity. In this work, our objective is to design and assess a structured framework to perform distribution matching on concepts described by continuous variables. We propose a framework that combines distribution matching and machine learning techniques. Using a training sample consisting of correct and incorrect correspondences between different terminologies, a match probability score is built. For each term, best candidates are returned and sorted in decreasing order using the probability given by the model. Searching 101 terms from Lille University Hospital among the same list of concepts in MIMIC-III, the model returned the correct match in the top 5 candidates for 96 of them (95%). Using this open-source framework with a top-k suggestions system could make the expert validation of terminologies alignment easier.

Keywords. ontology matching, health informatics, probability distribution, probability metrics

1. Introduction

The use of international laboratory terminologies (e.g. LOINC) inside hospital information systems is required to conduct data reuse analyses through inter-hospital databases. Well-known strategies from the ontology matching field of computer science have already been proposed, like string-based or language-based models [1], to standardize local terminologies toward an international reference.

Distribution matching is an *instance-based matching technique* [2] that performs terms matching based on the statistical similarity of their respective sets of instances. This technique is an extension of two-sample hypothesis testing to compare distributions.

Recent developments on *distribution matching* include a comprehensive evaluation of schema matching techniques [3], where a Wasserstein distance-based *distribution matching* algorithm [4] competes with state-of-the-art *schema-based matching*

[1] Corresponding Author, Martin COURTOIS, CHU Lille, INCLUDE, Institut Coeur-Poumon Boulevard du Professeur Jules Leclercq, 59000 Lille, France; E-mail: martin.courtois@chru-lille.fr, martin.courtois@protonmail.com.

techniques. To the best of our knowledge, *distribution matching* has only been applied once to healthcare terminologies alignment through the use of expert data preprocessing [5].

In this work, we design and assess a structured, reproducible framework to perform *distribution matching* in a real-world setting. This framework aims at aligning laboratory terminologies described by continuous variables without any required preprocessing while using *f*-divergences as similarity measures (e.g. the Hellinger distance).

2. Methods

2.1. The Distribution Matching Framework

In this section we propose a generic framework for distribution matching, applied on two experimental scenarios. This framework is based on a machine learning classification model which is trained on features describing the similarity of distributions. This model outputs the probability of a match for a given pair of terms from two distinct terminologies. Our framework consists in the following steps:

1. The set of all possible pairs is defined by the cartesian product of two terminologies. The equivalence (hence disjointness) of those pairs is defined manually.
2. For each pair, we compute distribution-based features from the measurements: (a) the Kolmogorov-Smirnov statistic, (b) the Hellinger distance, (c) the absolute difference of the means and (d) the absolute difference of the standard deviations.
3. We train and fine-tune a random forest classifier on the previous set of features.
4. For each pair, we use the model's probability of correct match to predict equivalence or disjointness. The model's predictive ability is then assessed using the following metrics: *Precision*, *Recall*, F_1 or *Precision-Recall AUC*. In practice, we are interested in producing mappings from one source to another. In this case, we can also use the *Mapping Score*, which measures the ability of the matching technique to provide a correct match in the top 5 ranked candidates.

2.2. Datasets and Scenarios

Our framework is assessed using data from the freely accessible database *MIMIC-III* and from the Lille University Hospital's laboratory terminology, the reuse of which for medical research purposes has been authorized by the CNIL in 2019 (reference number 2202081). All models were fine-tuned using 5-fold cross-validation on a training dataset and evaluated on a testing dataset. The training and testing datasets both contain the complete terminologies (i.e. the whole terms). Splitting is performed at random on the series of measurements according to a 70% (train) / 30% (test) ratio. We propose two experimental scenarios to assess the model's behaviour :

Case 1: We worked on a subset of 54 terms from the *MIMIC-III* database and 101 terms from the Lille University Hospital's laboratory terminology. These two sets share the same exact concept domain. A reference alignment was manually produced by a biologist with expertise in laboratory terminology. This alignment is composed of 5340 disjoint and 114 equivalent pairs for a total of 5454 pairs. In this use case, we computed the *Mapping Score* from the Lille University Hospital terminology to *MIMIC-III*, in addition to the *Precision-Recall AUC*. We used separate univariate logistic regressions to perform classification based on single features (e.g. KS statistic). A random forest classifier (our proposed model) was used to classify pairs through the combination of multiple features. *Case 2:* We propose a complementary analysis which is a positive control. In this use case, we tried to match *MIMIC-III* with itself on a subset of 195 terms. The reference alignment was produced automatically using terms' ids to identify the correct pairs.

3. Results

Our *distribution matching* framework is implemented in *Python* version 3.8 and is available under Apache-2.0 License at https://github.com/mcrts/dmatch. The developed package provides through a CLI the required tools to extract and prepare the data in order to train and evaluate a decision model between two given terminologies.

3.1. Experimental Results

Table 1 displays the evaluation metrics computed on the testing dataset from Lille University Hospital to *MIMIC-III* terminologies as part of use case 1. We provide *PrecisionRecall AUC* and the *Mapping Score* along with 95% confidence intervals for each model. Figure 1 shows the features' importance derived from the random forest classifier. Those are computed as the average sums of impurity decrease within each tree.

Table 1. Evaluation metrics of the decision models on the testing dataset (use case 1)

Model	*Precision-Recall AUC*	*Mapping Score*
Kolmogorov-Smirnov (KS) statistic	0.63 [0.53, 0.71]	0.97 [0.92, 0.99]
Hellinger distance	0.51 [0.41, 0.61]	0.81 [0.72, 0.88]
L^1 norm of the means	0.18 [0.13, 0.24]	0.82 [0.73, 0.89]
L^1 norm of the standard deviations	0.08 [0.05, 0.12]	0.46 [0.36, 0.56]
Our model (random forest)	0.67 [0.58, 0.77]	0.95 [0.89, 0.98]

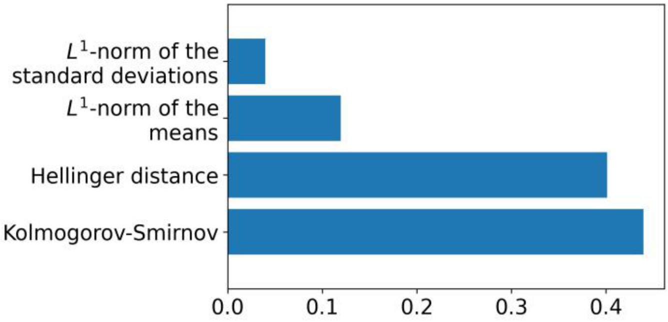


Figure 1. Random forest features' importance (use case 1)

3.2. Positive Control Use Case

When trained for the specific task of matching the *MIMIC-III* terminology with itself, the selected random forest model reaches a *Precision-Recall AUC* of 0.96[0.94,0.98] on testing pairs (use case 2). Table 2 shows the five model's suggestions with highest probability for the case of Monocytes cell count in cerebrospinal fluid (LOINC 26486-1).

Table 2. Top 5 candidates for Monocytes cells count in cerebrospinal fluid 26486-1

Laboratory Term in MIMIC-III	Nature	Probability
51120 — Monocytes — Ascites — 26488-7	False	0.988
51355 — Monocytes — CSF — 26486-1	True	0.912
50801 — Alveolar-arterial Gradient — Blood — 19991-9	False	0.000
51130 — Absolute CD3 Count — Blood — 8124-0	False	0.000
51332 — Absolute CD8 Count — Blood — 8138-0	False	0.000

4. Discussion

In this work, we applied distribution analysis to match laboratory terminologies between hospitals. The objective was to explore the use of distribution-based similarity measures for terminology matching, implement and benchmark this technique against uncurated laboratory data from the *MIMIC-III* database and the Lille University Hospital. The selected model was able to give the correct correspondence among the 5 best candidates for 95% of the 101 terms considered. As illustrated by the overall *PrecisionRecall AUC* and features' importance, distribution-based similarity measures such as the KS statistic and the Hellinger distance strongly improve the performance of the decision model compared to the absolute difference of the means or standard deviations.

A second use case (positive control) consisted in matching the *MIMIC-III* dataset against itself. The model built as part of this use case gave near perfect results which illustrates the general feasibility of our framework.

4.1. Methodological Issues

As opposed to conventional language-based matching techniques, *distribution matching* does not rely on the quality and richness of terminologies. Indeed, it showed to be resilient to data anomaly when tested on uncurated datasets. However, in its current state, our framework remains sensitive to mismatching unit systems between data sources.

In practice, the Hellinger distance relies on kernel density estimates which are sensitive to ill-behaving data sample and requires cpu intensive numerical integration. In spite of this limitations, our model still remains accurate thanks to the combination of other distribution-based features using ensemble learning. At last, our framework yet supports only univariate distribution of continuous variables.

4.2. Perspectives

In this work, we focused on the Kolmogorov-Smirnov statistic and the Hellinger distance. Other distribution-based similarity measures can also be used [6], especially the Integral Probability Metrics for which efficient computation techniques exist [7].

To further evaluate our *distribution matching* framework, we intend to benchmark it against regular language-based technique using only publicly available data such as AmsterdamUMCdb [8]. A composite model combining language and distribution analysis could then be trained to reach better performances. In particular, we believe that using top-*k* suggestions could make the expert validation of terminologies alignment easier. Through the alignment of *MIMIC-III* terminology with itself, our framework can be used for assessing the quality and consistency of a single terminology. Thus, we believe that such a tool could be used to detect terminology's anomalies, e.g. the modification of the identifier of a concept over time (especially for local terminologies).

As part of the operational setting of hospital data processing, a single concept usually has different identifiers in a single terminology for each laboratory or production site. This issue could also be addressed by the proposed framework.

5. Conclusion

In this study, we proposed a framework that combines *distribution matching* and machine learning techniques for terminology matching in a clinical setting. We trained and evaluated an algorithm on two scenarios and identified operational use cases. Finally, we provided a frame of reference that will pave the way for future improvements.

References

[1] Khan AN, Griffith SP, Moore C, Russell D, Rosario AC, Bertolli J. Standardizing laboratory data by mapping to LOINC. Journal of the American Medical Informatics Association. 2006;13(3):353-5.

[2] Euzenat J, Shvaiko P. Ontology matching. 2nd ed. Heidelberg (DE): Springer-Verlag; 2013.

[3] Koutras C, Siachamis G, Ionescu A, Psarakis K, Brons J, Fragkoulis M, et al. Valentine: Evaluating Matching Techniques for Dataset Discovery. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE); 2021. p. 468-79.

[4] Zhang M, Hadjieleftheriou M, Ooi BC, Procopiuc CM, Srivastava D. Automatic Discovery of Attributes in Relational Databases. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. SIGMOD '11. New York, NY, USA: Association for Computing Machinery; 2011. p.109–120. Available from: https://doi.org/10.1145/1989323.1989336.

[5] Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. In: AMIA Annual Symposium Proceedings. vol. 2011. American Medical Informatics Association; 2011. p. 392.

[6] Cha SH. Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. Int J Math Model Meth Appl Sci. 2007 01;1.

[7] Sriperumbudur BK, Gretton A, Fukumizu K, Scholkopf B, Lanckriet G. Hilbert space embeddings and¨ metrics on probability measures. The Journal of Machine Learning Research. 2010;11:1517-61.

[8] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration. Critical Care Medicine. 2021 Feb;Publish Ahead of Print. Available from: https://doi.org/10.1097/ccm.0000000000004916.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210824

A Methodology for an Auto-Generated and Auto-Maintained HL7 FHIR OWL Ontology for Health Data Management

Vassilis KILINTZIS[a,1], Vasileios C. ALEXANDROPOULOS[a],
Nikolaos BEREDIMAS[a] and Nicos MAGLAVERAS[a]
[a] *Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies, Medical School, Aristotle University of Thessaloniki, Greece*

Abstract. The process of maintenance of an underlying semantic model that supports data management and addresses the interoperability challenges in the domain of telemedicine and integrated care is not a trivial task when performed manually. We present a methodology that leverages the provided serializations of the Health Level Seven International (HL7) Fast Health Interoperability Resources (FHIR) specification to generate a fully functional OWL ontology along with the semantic provisions for maintaining functionality upon future changes of the standard. The developed software makes a complete conversion of the HL7 FHIR Resources along with their properties and their semantics and restrictions. It covers all FHIR data types (primitive and complex) along with all defined resource types. It can operate to build an ontology from scratch or to update an existing ontology, providing the semantics that are needed, to preserve information described using previous versions of the standard. All the results based on the latest version of HL7 FHIR as a Web Ontology Language (OWL-DL) ontology are publicly available for reuse and extension.

Keywords. HL7 FHIR, OWL, RDF

1. Introduction

As population health management (PHM) becomes the new best practice of healthcare, a new information technology infrastructure is needed to facilitate this care delivery model. Within this infrastructure electronic health records (EHRs) are necessary but not sufficient. Interoperability among health IT systems and with other data sources is crucial to PHM but is still far from being achieved. Interoperability advocates are promoting the development of the Fast Health Interoperability Resources (FHIR) standard and the use of open application programming interfaces (APIs) [1].

Semantic technologies and Linked Data principles [2] are in the heart of the solution that entails multiple sources of data and information along with multiple access points, a strong temporal aspect, as well as different computational workflows.

FAIR Principles emphasize on the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention [3]. To

[1] Corresponding Author, Vassilis Kilintzis; E-mail: billyk@med.auth.gr.

that end, the use of standards and definition of detailed semantics, as early as possible, in the process of domain definition is critical.

Several public or commercial repositories are modeling knowledge using semantic web technologies. One of the easiest to use and freely accessible repository of health-related ontologies is Bioportal of the National Center for Biomedical Ontology [4]. It incorporates search and representation mechanisms for several health ontologies and terminologies such as SNOMED Clinical Terms, International Classification of Diseases (ICD) and Logical Observation Identifier Names and Codes (LOINC) among others. There is additional work being done in the process of transforming OWL ontologies into FHIR terminology resources as presented in [5]. The paper presents the challenges in the transformation a detailed overview of the mapping between OWL and FHIR.

In our previous work we have presented, initially, a representation as an ontology of the HL7 FHIR primitive and complex data types [6], then, a platform to support integrated care built upon linked data principles based on an ontology representation of HL7 FHIR[7] along with a methodology of maintenance of this ontology, aiming to keep up with the evolving standard and at the same time retain backwards compatibility with the software built upon previous versions[8].

In this paper we present the methodology used to develop a FHIR ontology generator software used for automatically transforming the FHIR specification resource files into an OWL ontology. The resulting ontology is ready to support the data management of EHR and PHM data along with the provision for managing future evolution of the standard. It includes all the FHIR defined data types and resources and, can be further expanded or restricted using owl axioms to adhere to a specific domain concepts or restriction on data. The ontology is shared publicly and aims on reusability by providing access to a very broad and componentized data model, helping researchers and adopters to overcome the triviality of re-implementing the base health-domain model each time.

2. Methods

The FHIR standard defines three types of data structures. Primitive data types, which are simple values, like string or decimal, complex data types, which are reusable collections of primitive types or other complex types, like *ContactDetails* or *HumanName*, and resources, which represent specific parts of the medical process and contain the data required to define them, like *Observation* or *AllergyIntolerance*.

In our ontology, each of the three data structure types are defined as an *owl:Class*. While primitive types could potentially be defined as *owl:Datatype*, *owl:Class* is used due to the absence of software support for custom data types. The properties of complex types and resources are mapped to *owl:ObjectProperty* in our ontology. Further details on the design decisions of the ontology can be found in [6] and [8].

2.1. FHIR Ontology Generator

FHIR provides various metamodels of the data structures defined as part of it[2]. For our needs, we use a JSON document defining the complex types and resources as instances

[2] https://www.hl7.org/fhir/downloads.htm

of *StructureDefinition* which is a meta-resource defined by FHIR to facilitate the exchange of custom resources. Primitive types are for the most part are “set in stone”, so they are not created as a part of the auto generator. An existing implementation is instead added to the created ontology. The simple mappings are presented in Table 1.

Property “kind” can have one of four values “logical”, “resource”, “complex-type”, “primitive-type”. Values “logical” and “resource” have similar semantics and are treated the same. Value “primitive-type” is ignored as discussed previously.

Property “name”, is used as the URI suffix of the resulting *owl:Class*.

The last property “differential.element” of *StructureDefinition*, is an array of *ElementDefinition*. *ElementDefinition* is a meta-resource that defines a single element of the modelled resource (e.g. the semantics of “*Observation.referenceRange.type*” element of the *“Observation”* resource are described as an *ElementDefinition*). Each *ElementDefinition* in the array is mapped to an *owl:ObjectProperty* assigned via *rdfs:domain* to the specific *owl:Class*. The semantics accompanying the specific *ElementDefinition* are defined at the corresponding *owl:Class*.

The property “path” is used to identify the *rdfs:domain* of the *owl:ObjectProperty.* In FHIR, resources often contain elements that need to be grouped together in a single substructure. When a substructure, like this, is defined only in the context of a single resource, it is modeled as an extension of the complex type *BackboneElement.* Depending on the value of “path” the *rdfs:domain* is either an existing *owl:Class* (corresponding to either a resource or of a complex-type), or a new *owl:Class* (*rdfs:subClassOf FHIRct:BackboneElement*). Property “path” is also used to generate the *rdfs:label* of the *owl:ObjectProperty*.

Property “type” includes one or many substructures that determine the allowed values for this element. There are two cases. In one case, when the allowed values can be a simple or complex type, the value becomes the *rdfs:range* of the property. In the other case the allowed value is a reference to a resource so the allowed types become the *rdfs:range* and the custom annotation *takesReference* is added to the *owl:ObjectProperty* to allow easier handling in applications using it.

Table 1. FHIR to OWL mapping

FHIR StructureDefinition	FHIR ElementDefinition	OWL Namespace:Local Name
name	path	rdfs:label
version		owl:version
description	definition	rdfs:comment
baseDefinition		rdfs:subClassOf
differential.element		owl:ObjectProperty
	min, max	owl:minCardinality, owl:maxCardinality, owl:cardinality
	isSummary	pt:isSummary
	isModifier	pt:isModifier

pt is the namespace: *http://lomi.med.auth.gr/ontologies/FHIRPrimitiveTypes#*

2.2. Integrating version changes to existing ontologies

The ontology generator described in 2.1 is sufficient when building an ontology based on the latest version of FHIR from the ground up. When a new version of the standard is introduced and since ontologies of previous versions could be already used, updates in the resources or complex data types can cause problems. A methodology for managing these possible problems, aiming to maintain existing data and their semantics, even if they were defined by previous versions of the standard, using the expressivity provided by OWL is presented in [8]. The presented generator is using a FHIR-provided JSON document with the changes between versions to tackle this issue automatically. The useful types of changes documented are the following:

- Addition or deletion of elements, or name change to elements
- Changes to minimum and maximum cardinality of a property
- Changes to allowed types of a property

The JSON document contains an object for each FHIR data structure with a property "status". In the case of "changed", an object "elements" is provided that has a key for each of the changed properties of this particular data structure. The value of each key in "elements" is an object that may contain one or more definitions. The possible combinations are shown in Table 2.

Table 2. Handling of changes between versions

Status	Property combination	Changes to owl:Class	Changes to owl:ObjectProperty
deleted		Add owl:deprecated	Add owl:deprecated
changed	old-min – new-min	Change owl:minCardinality	
	old-max – new-max	Change owl:maxCardinality	
	removed-types		Remove from rdfs:range
	added-types		Add to rdfs:range
new		Create new Class	Create new ObjectProperty

3. Results

The generator is implemented as a Java app to take advantage of the Apache Jena API for building the ontology. Upon execution the corresponding ontology is created as an RDF/turtle file with the following base URIs:

http://lomi.med.auth.gr/ontologies/FHIRComplexTypes
http://lomi.med.auth.gr/ontologies/FHIRPrimitiveTypes
http://lomi.med.auth.gr/ontologies/FHIRResources

The procedure of the ontology update was tested using FHIR Release 3 (STU) and the current version FHIR Release #4.

The OWL-DL ontology automatically produced by the generator from FHIR v4.0.1: R4, is available for review and reuse in http://lomi.med.auth.gr/ontologies/FHIR

4. Discussion

While a semantic data model represented as an OWL ontology is an obvious choice to provide semantics to the modeled entities, and efforts of transforming relational databases to ontologies and vice-versa are widespread [9], using an ontology to describe, store and exchange actual health related data is not as common. Having HL7 FHIR as the de facto standard for exchanging health data, a system based on an ontology, derived from FHIR, could minimize data mappings and transformations, semantically enrich the managed health data to tackle interoperability obstacles and enhance findability, as proposed by FAIR principles. Such a system presented in [8] requires maintenance of the FHIR based ontology and this process requires a lot of effort and attention.

In this paper we have presented the FHIR ontology Generator software that is capable of automatically transforming the provided by the standard definition files to an OWL-DL ontology. The resulting ontology includes all the semantics needed to support EHR data storage with data validation (i.e., accepted values, cardinality of values) along with support for automatic application into the ontology of future changes in the standard, maintaining intra-model semantics for backwards compatibility.

FHIR ontology Generator software can reduce the interoperability gap exhibited among health record systems and offer a model for PHM by providing either a solid data model base for specific domain implementations or a reference model to map existing deployed models. In the first option, the data model base, generated from the latest version of HL7 FHIR standard, can be restricted to accept specific domain entities with additional semantics, while in the second options, mapping to the model can be used, as a first step in the pipeline, to export FHIR resources from non-FHIR based EHR system.

Next steps include the integration, in the software, of the transformation rules that aim at the in-FHIR defined terminologies and bindings, as well as the development of a new software for the export of ontology instances as JSON FHIR resources.

Acknowledgements

Research was partially supported by the EU-WELMO project (project number 210510516).

References

[1] Watson Health, Population health management beyond the EHR: Part 2 - Watson Health Perspectives, *IBM Watson Health Perspectives*, 2017. https://www.ibm.com/blogs/watson-health/population-health-management-beyond-ehr-part-2/ (accessed Apr. 21, 2021).

[2] Berners-Lee T, Linked Data, 2011. https://www.w3.org/DesignIssues/LinkedData.html (accessed Feb. 18, 2015).

[3] FAIR Principles, https://www.go-fair.org/fair-principles/ (accessed Jul. 1, 2021)

[4] Whetzel PL, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research. 2011 Jun 14;39(suppl_2):W541-5. .

[5] Metke-Jimenez A, Lawley M, Hansen D. FHIR OWL: Transforming OWL ontologies into FHIR terminology resources. InAMIA Annual Symposium Proceedings 2019 (Vol. 2019, p. 664). American Medical Informatics Association..

[6] Beredimas N, Kilintzis V, Chouvarda I, Maglaveras N. A reusable ontology for primitive and complex HL7 FHIR data types. In2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2015 Aug 25 (pp. 2547-2550). IEEE.

[7] Kilintzis V, Chouvarda I, Beredimas N, Natsiavas P, Maglaveras N. Supporting integrated care with a flexible data management framework built upon Linked Data, HL7 FHIR and ontologies. Journal of biomedical informatics. 2019 Jun 1;94:103179.

[8] Kilintzis V, Kosvyra A, Beredimas N, Natsiavas P, Maglaveras N, Chouvarda I. A sustainable HL7 FHIR based ontology for PHR data. In2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2019 Jul 23 (pp. 5700-5703). IEEE.

[9] Spanos DE, Stavrou P, Mitrou N. Bringing relational databases into the semantic web: A survey. Semantic Web. 2012 Jan 1;3(2):169-209.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210825

Assessing Resolvability and Consistency in OBO Foundry Ontologies: Pilot Study

Shuxin ZHANG[a,1], Nirupama BENIS[a] and Ronald CORNET[a]
[a] *Department of Medical Informatics, Amsterdam University Medical Center, Amsterdam, The Netherlands*

Abstract. Ontologies listed in the OBO Foundry are often regarded as reliable choices to be reused but ontology interoperability of them remains unknown. This study evaluated the resolvability of URIs and consistency of axioms in the OBO Foundry library, BFO ontology, and CIDO ontology. All had nonresolvable URIs, but the OBO library and the CIDO had additional interoperability issues regarding the use of incorrect prefixes, mixing up with ontologies, and inconsistency in the use of property. These detected issues reflected the real-world common problems that were not significant from human beings' point of view but hindered the machine-processability of ontologies. The assessment performed in this study was automated and enables scale-up against more metrics over more ontologies, which remains future work.

Keywords. Ontologies, Interoperability, Assessment, OBO Foundry

1. Introduction

Ontologies, as means to formalize concepts and relations that represent entities and their relations in a specific domain of the world, support health data in being Findable Accessible, Interoperable, and Reusable (i.e., FAIR). Given a large number of ontologies developed [8][7], ontology interoperability is needed to prevent misunderstanding and isolation between different resources [1]. The Open Biological and Biomedical Ontologies (OBO) Foundry [8] is an ontology library making registered ontologies more findable and reusable. These ontologies are reviewed and listed in the OBO Foundry library, and in practice, they are often regarded as "good" choice to be reused. However, "being commonly-used" or "being of good reputation"[2] are not the golden standards but subjective preferences, which can make people unconsciously unaware of detectable pitfalls that should not be further spread over by those "good" ontologies. In this pilot study, we examined the resolvability of URIs and consistency of RDF triples of published ontologies in the OBO Foundry library to explore the real-world interoperability issues.

[1] Corresponding Author.

[2] http://purl.org/ist/interoperability-key-point#commonly-used-vocabulary

2. Methods

We applied the assessment approach from [10] which evaluates the interoperability of existing linked dataset in RDF. Five metrics (see Table 1) were selected because:

- they are objective with minimal human involvement;
- their testing can be automated;
- they reflect four different dimensions[3].

Table 1. List of metrics implemented in this study and their reflected dimensions.

Dimension	Metric	Interpretation
Availability	Resolvability of URIs	Upon request of a URI term, check whether any information is provided as result.
Representational-consistency	Reuse of existing terms	Detect the use of existing terms.
Understandability	Use of Human-readable Labelling	Detect the use of human-readable annotations.
Consistency	Misplaced classes or properties	If classes are correctly used as objects in rdf:type triples. If properties are correctly used as predicates in triples.
	Misused Properties of the type owl:DatatypeProperty or owl:ObjectProperty	If objects of properties as type “owl:DatatypeProperty” are literal. If objects of properties as type owl:ObjectProperty are URIs.

Three RDF datasets were evaluated (see Table 2) as representatives of their types. BFO and CIDO, which are represented in the Web Ontology Language (OWL)[4], are regarded as RDF datasets in this study. BFO has been manually reviewed by experts in the OBO Foundry community while reviewing CIDO is not completed yet. They are commonly-used and can serve as the starting point for assessing ontologies in OBO Foundry.

Table 2. List of evaluated RDF datasets.

Dataset	Type	Description
OBO Library	Metadataset	A dataset which lists current OBO ontologies with their meta information, including activity status, access URI, theme, and etc.
BFO	Upper Ontology	An upper-level ontology in support of domain ontologies developed for scientific research within the framework of OBO Foundry.
CIDO	Domain Ontology	A biomedical ontology in the area of coronavirus infectious disease.

Unique URIs were extracted from RDF datasets to check resolvability, which means that an HTTP request for these URIs can provides us with other resources. Results were described by HTTP status code. Diagram of workflow can be found at a persistent URL[5].

RDF triples were extracted and then divided into triples with or without the property rdf:type. The types of objects in rdf:type triples were checked if they were of type owl:Class. Predicates in other triples were checked if they were of any property type (e.g., rdfs:Property). After that, we examined if these classes and properties were processable by machines, which checks 1) if any content can be automatically retrieved via either parsing through a parser or querying through a SPARQL wrapper and 2) if retrieved content contains given resource of that URI. For the predicates whose types

[3] https://purl.org/iqd

[4] https://www.w3.org/TR/owl2-overview/

[5] https://purl.org//report/workflow

were specified as either owl:DatatypeProperty (whose objects should be literal) or owl:ObjectProperty (whose objects should be URI), their objects in triples were checked. Any URI, class, predicate, or triple that was assessed but failed against metrics was regarded as a failure case, which served as the unit for analysis. We utilized the parser and SPARQL wrapper developed in *rdflib*[6] package. The SPARQL endpoints were Ontobee[7] and BioPortal[8]. Implementation scripts can be found at GitHub[9].

3. Results

Table 3 describes the number of failure cases detected in the OBO library dataset, BFO and CIDO against metrics. The OBO library dataset has failures in unavailability of 38 (out of 1,067) URIs, unretrievability of 5 (out of 23) predicates, and 8 (out of 14) misused properties of owl:ObjectProperty, and these failures also occurred in CIDO, while the BFO dataset has failures only in unavailability of 70 URIs (out of 156). In terms of Understandability, all datasets applied human-readable labelling, including rdfs:label and dct:description.

All test datasets have problems in resolvability of URIs and some of them stemmed from the same resource. In the OBO library dataset, all terms with the prefix "http://obofoundry.github.io/vocabulary/" were not found (HTTP 404). Some URIs are those referring to deprecated ontologies, e.g., <http://purl.obolibrary.org/obo/epo.owl> In the BFO, all URIs with the prefix "http://purl.obolibrary.org/obo/bfo/axiom/" are not found (HTTP 404), which amount to 67 (out of 70). In the OBO dataset, a property <http://purl.org/dc/terms/1.1/theme> was used but it does not exist, though that URI is still resolvable to DCMI Metadata Terms (DCT) ontology. Through query via a SPARQL endpoint, this error was detected as that query was performed by extract pattern matching. <http://www.w3.org/ns/dcat#theme> from Data Catalog Vocabulary (DCAT), however, exists. So it is important to distinguish terms between DCT and DCAT, alike but different. Besides, DCT maintains two namespaces: "http://purl.org/dc/elements/1.1/" and "http://purl.org/dc/terms/" but not "http://purl.org/dc/terms/1.1/". So the URI <http://purl.org/dc/terms/1.1/license> in the OBO library dataset could not resolve to any content regarding license and was not queriable in SPARQL endpoint. Therefore, we should be aware of correct use of DCT namespaces, though such mistake still can guide you towards DCT resources but it is not processable by machines. Eight failed properties of owl:ObjectProperty in the OBO dataset and ten failure in CIDO are listed[10] with number of involved RDF triples. In all of these failed triples, we found that all objects were the string version of an URI instead of the Notation 3 format[11]. Below is an example:

```
<http://purl.obolibrary.org/obo/obi>
<http://usefulinc.com/ns/doap#bug-database>
"http://purl.obolibrary.org/obo/obi/tracker"
```

[6] https://github.com/RDFLib/rdflib

[7] http://www.ontobee.org/sparql

[8] http://sparql.bioontology.org/

[9] https://github.com/sxzhang1201/Interoperable-Supportive-Tool

[10] https://purl.org/obo_library_assess/object_property

[11] https://www.w3.org/TeamSubmission/n3/

Table 3. Number of failure cases against metrics for OBO library and BFO.

Metrics	OBO Library Dataset		BFO		CIDO	
	# Total	**# Failure (%)**	**# Total**	**# Failure (%)**	**# Total**	**# Failure (%)**
Resolvability of URIs						
- Available URIs	1067	38 (4%)	156	70 (45%)	9598	2183 (23%)
Reuse of existing terms						
- URI of a class resolving to content concerning that class	4	0	19	0	2249	117 (5%)
- URI of a predicate resolving to content concerning that predicate	23	5 (22%)	26	0	164	33 (20%)
Misplaced classes or properties						
- Classes incorrectly used as properties	4	0	19	0	2132	0
- Properties incorrectly used as classes	23	0	26	0	131	0
Misused Properties of the type owl:DatatypeProperty or owl:ObjectProperty						
- Properties of owl:DatatypetProperty	1	0	5	0	40	4 (10%)
- Properties of owl:ObjectProperty	14	8 (57%)	4	0	13	10 (77%)
Human-readable Labelling						
- Human-readable annotations	609	0	50	0	9402	0

4. Discussion

In this study, we found that well-used ontologies from a reliable platform contained errors, including non-resolvability of URIs, use of incorrect prefixes, mixing up with ontologies, and inconsistency in the use of property. Both OBO library and BFO maintain their own vocabularies, all of which, however, are not resolvable. It is probably ascribed to authorization issues and further information is needed by reaching out to their authors. URIs referring to deprecated ontologies were not but should be resolvable along with version information so that those still using outdated ones are able to find related update activity and reach out to the updated ones. Many researchers have performed quality assessment of ontologies. Burton-Jones et al.[2] proposed a suite of metrics, i.e., Syntactic, Semantic, Pragmatic, and Social, to evaluate the usefulness of ontologies found in the DARPA Agent Markup Language (DAML) library. The only metric relevant to Consistency in [2] measures the proportion of inconsistent classes and properties but does not clarify how such inconsistency could be detected. Duque-Ramos et al.[4] adapted a Software Engineering standard, Software product Quality Requirements and Evaluation (SQuaRE) to develop a framework for ontology evaluation. Fourteen metrics were defined to assess the quality of ten ontologies of "units of measurements" and "cell types". These metrics were measured in an automated manner but focused on "demographics", for example, measuring the number of attributes per class, and the mean number of direct subclasses. He et al.[5] proposed an "eXtensible Ontology Development" strategy and four associated principles (i.e., ontology reuse, ontology semantic alignment) to provide high-level guideline for ontology development. Our study instead focused on a relatively lower level of quality assessment enabling resolvability and consistency checking in an automated and expectedly complete fashion.Our study employed an automated approach to assess a set of metrics reflecting different dimensions of ontology interoperability. Such automation enables to evaluate more datasets in an objective way. However, there are more quality metrics in [10] that are not tested. The tool implemented is capable of a limited number of metrics but incorporation with other existing tools, e.g., Luzzu[3] and RDFUnit[6], can support the expansion of quality assessment. An integrated assessment approach performed by Sanju et al.[9] is also promising to detect additional interoperability problems but inconsistency of performance among different assessment tools should be addressed. True machine readability of ontologies, concepts and classes is key to supporting reasoning over data and establishing FAIR linkable data. Consequently, the quality of such ontologies should be maximal, hence quality assessment should be applied, and should be facilitated. Our approach contributes to quality assessment, and the developed tool automates such assessment. In the future, with more metrics incorporated, more ontologies should be assessed to capture a comprehensive view of common interoperability problems in existing well-used ontologies of a specific domain, for example, ontologies concerning COVID-19.

5. Conclusions

Even established, well-used ontologies aren't free of errors that can be automatically detected. We have developed tooling that helps to detect and resolve errors. Further work and research are needed to detect more types of errors over more ontologies.

References

[1] Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. Journal of biomedical informatics. 2018 Apr 1;80:1-3.

[2] Burton-Jones A, Storey VC, Sugumaran V, Ahluwalia P. A semiotic metrics suite for assessing the quality of ontologies. Data & Knowledge Engineering. 2005 Oct 1;55(1):84-102.

[3] Debattista J, Auer S, Lange C. Luzzu—a methodology and framework for linked data quality assessment. Journal of Data and Information Quality (JDIQ). 2016 Oct 25;8(1):1-32..

[4] Duque-Ramos A, et al. OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. Journal of research and practice in information technology. 2011 May;43(2):159-76.

[5] He Y, et al. The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. Journal of biomedical semantics. 2018 Dec;9(1):1-0.

[6] Kontokostas D, Westphal P, et al. Test-driven evaluation of linked data quality. InProceedings of the 23rd international conference on World Wide Web 2014 Apr 7 (pp. 747-758)..

[7] Corbin-Lickfett KA, et al. The HSV-1 ICP27 RGG box specifically binds flexible, GC-rich sequences but not G-quartet structures. Nucleic acids research. 2009 Nov 1;37(21):7290-301..

[8] Smith B, Ashburner M, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007 Nov;25(11):1251-5.

[9] Tiwari S, Abraham A. Semantic assessment of smart healthcare ontology. International Journal of Web Information Systems. 2020 Jul 31.

[10] Zhang S, Benis N, De Keizer N, Cornet R. An Approach for Interoperability Assessment of RDF Data, *Submitt. to Semant. Web J.* (2021).

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210826

Intelligent Integrative Platform for Sharing Heterogenuous Stem Cell Research Data

Kirill BORZIAK[a,1], Irena PARVANOVA[a] and Joseph FINKELSTEIN[a]
[a] *Icahn School of Medicine at Mount Sinai, New York, New York, USA*

Abstract. Recent studies demonstrated that comparative analysis of stem cell research data sets originating from multiple studies can produce new information and help with hypotheses generation. Effective approaches for incorporating multiple diverse heterogeneous data sets collected from stem cell projects into a harmonized project-based framework have been lacking. Here, we provide an intelligent informatics solution for integrating comprehensive characterizations of stem cells with research subject and project outcome information. Our platform is the first to seamlessly integrate information from iPSCs and cancer stem cell research into a single platform, using a multi-modular common data element framework. Heterogeneous data is validated using predefined ontologies and stored in a relational database, to ensure data quality and ease of access. Testing was performed using 103 published, publicly-available iPSC and cancer stem cell projects conducted in clinical, preclinical and in vitro evaluations. We validated the robustness of the platform, by seamlessly harmonizing diverse data elements, and demonstrated its potential for knowledge generation through the aggregation and harmonization of data. Future aims of this project include increasing the database size using crowdsourcing and natural language processing functionalities. The platform is publicly available at https://remedy.mssm.edu/.

Keywords. Common data elements, induced pluripotent stem cells, cancer stem cells.

1. Introduction

Stem cells were first described in 1961 by James Till and Ernest McCulloch [1]. Today, stem cell research has dramatically transformed and advanced the field of regenerative medicine. Due to the large number of published stem cell research studies, researchers aim to collect, store, and centralize the gathered data. In previous publications by our team, we have developed and tested Regenerative Medicine Data Repository (ReMeDy) platform, allowing collection and sharing of *in vitro* findings and pre-clinical/ clinical trial outcomes [2, 3]. Currently, our platform contains 103 stem cells research papers, included in the PubMed database. Each featured project can be accessed across the framework by utilization of user-friendly tools and API platforms, due to the use multi-modal flexible common data elements (CDE) framework, which permits cross-studies comparison and collaboration.

[1] Corresponding Author, Kirill Borziak, PhD, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA; E-mail: Kirill.Borziak@mountsinai.org.

2. Methods

2.1. Database architecture and web interface

Our platform, **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) [1], is an implementation of the Signature Commons (https://github.com/MaayanLab/signature-commons), which is a BD2K-LINCS DCIC platform [2], installed through Docker and designed to store and search diverse metadata in an agile and flexible manner [4]. The ReMeDy platform was installed using the default instructions on a Linux server. It contains six repositories: controller, data-api, metadata-api, proxy, schema, and ui.

The various validation, visualization, and user interface schema were ingested through the Application programming interface (API) functionality. Specifically, we developed counting schemas based on the CDE framework, which aim to provide additional counting and filtering functionality to the search results page. The schemas, formatted in JSON, were generated and ingested using a custom Python script. To improve the utility of the API, we developed an upload interface, which automated the ingestion process. The upload interface was developed using ReactJS and Spring Boot. The interface allows for uploading and ingestion of CDE templates without command line interface, while maintaining the validation features.

2.2. Literature search and data abstraction

To test the ability of ReMeDy to handle heterogeneous stem cell data, we selected a set of 103 iPSC and CSC original research publications, using a randomized process from Google Scholar and PubMed search results for "iPSC" and "cancer stem cells", respectively. The randomized selection process was designed to ensure the inclusion of the full range of stem cell research. Further, we ensured the inclusion of in vitro, pre-clinical/animal model, and clinical trials of iPSC and CSC publications.

Following the selection of our publication set, the data from the publications was abstracted into the multi-modular Common Data Elements (CDE) framework [2, 4]. The abstraction process was conducted manually by trained abstractors with experience in cancer, regenerative medicine, and stem cell research. The majority of abstracted CDE values were defined either by permissible value sets or by ontologies. CDEs which are not amenable to being extracted as specific values, such as outcomes and findings descriptions, were recorded as short statements in free-text value fields. Further, a template was created for each cell line, individual, or grouped study subjects. The templates were then submitted to the upload interface utility, converted to JSON, ingested, and validated trough the API [5].

3. Results

The ReMeDy platform is a user-friendly database, which contains comprehensive and detailed information from stem cell research publications. The focus of the current iteration of ReMeDy is to seamlessly integrate induced pluripotent stem cell (iPSC) and cancer stem cell (CSC) projects. ReMeDy is currently freely accessible with no registration requirements.

3.1. The ReMeDy platform

The ReMeDy platform takes advantage of a relational database for data storage, such as PostgreSQL, which is implemented in our platform, excel at storing and searching structured data through organizing data within a well-defined schema (Figure 1). With the aim to conform to the FAIR guidelines (Findable, Accessible, Interoperable, and Reusable), our requirements for well-defined schema, validation against reference ontologies, ease and specificity of searching, and the ability to update data without compromising its integrity drove us to select it over a NoSQL approach. Further, indexing of the data enables for very fast searching of any attribute of the metadata without major slowdowns as the size of the tables expand. Our stringent metadata validation process includes strict definitions of key value pairs, the proper formatting of the values, and specification of required elements.

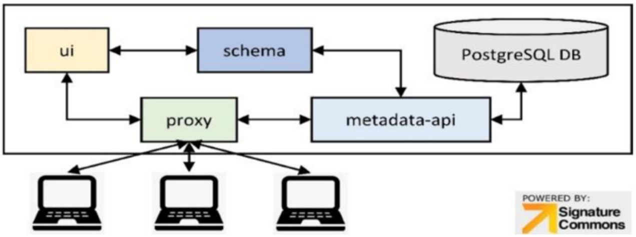


Figure 1. ReMeDy platform architecture displaying the interconnection of the Signature Commons packages.

3.2. Multi-modular CDE framework

In order to promote data harmonization and to facilitate data abstraction, we developed the multi-modular CDE framework. Our aim was to capture all the various facets of information related to iPSC and CSC projects. Previously standardized frameworks for characterization of stem cells, such as the Minimum Information About a Cellular Assay for Regenerative Medicine [6], do not cover the full range of information available from published projects and are limited to stem cell features and assays used to derive them. Our multi-modular CDE framework addresses these deficiencies by using a scoping review approach for defining relevant stem cell characteristic-related CDEs [7]. The resulting framework consists of 5 modules: Project, Stem Cell Characteristics, In-depth Characterization, Research System, and Outcomes / Findings (Figure 2).

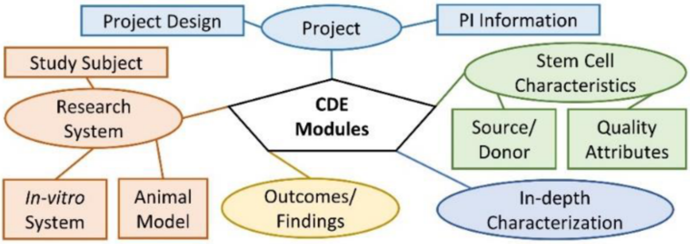


Figure 2. Schematic of the multi-modular CDE template, highlighting the modules and their CDE content.

The Project module CDEs capture general project information, such as PI contact information, funding information, publication information, and project design. The Stem Cell Characteristics module is designed to capture information about the stem cell products under investigation. The In-depth Characterization module contains CDEs related to different assays that can be used to characterize the stem cell, such as transcriptomic profiling, clonal capacity, or genetic stability. The Research System module CDEs characterize the study patients, animal models, and/or *in-vitro* cell lines. Finally, the Outcomes / Findings module CDEs describe the outcomes of clinical studies and findings from pre-clinical studies. Since not all CDEs are required for all studies, our modular organization provides a flexible approach for comparisons across studies.

3.3. Data accessibility, visualization and sharing

The ReMeDy site provides easy access to the various functionalities, such as search functionality, visualization tools, and API. It allows a search by CDE name or CDE value. Further, implemented filtering schemas allow users to incremental refinement of their search queries, and provide statistical information on the distribution of CDE values among the ReMeDy projects. ReMeDy also allows researchers to download the abstracter data directly through the API with the aim of promoting easy access, community sharing, and collaboration to advance stem cell research.

3.4. ReMeDy feasibility testing

To test the functionality and feasibility of our platform, we used 103 published clinical, pre-clinical, and *in vitro* iPSC and CSC studies. We abstracted on average 76 CDEs per study of total of 841 CDEs comprising the multi-modular framework. ReMeDy's feasibility was demonstrated by diversity of publications from the US, China, Japan, and Italy, amongst others. Abstraction of a wide range of source cell materials was tested (skin, blood, bone marrow, and others). Pre-clinical studies included studies in mice, rats, pigs, and rhesus macaques. We were able to abstract 15 different disease conditions, including cancer, heart disease, sclerosis, spinal cord injury, and others (Figure 3).

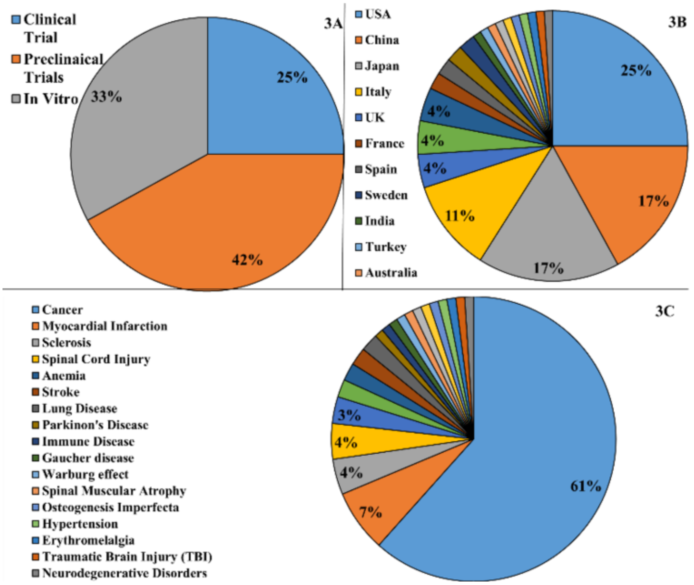


Figure 3. Distribution of projects in the ReMeDy platform across A. Project type; B. Country, conducting the research; C. Cancer type.

4. Discussion

The expanding field of stem cell research in both regenerative and cancer medicine requires the creation of a flexible and agile repository for data aggregation, storage, visualization, and sharing. To promote this effort, we have adapted the Regenerative Medicine Data Repository (ReMeDy) platform and the multi-modular CDE framework for use with both iPSC and CSC projects. A primary advantages of ReMeDy is its organized multi-modular framework, which harmoniously captures both iPSC and CSC research project information in a standardized format and provides effortless visualization. The platform was tested by uploading 103 clinical, preclinical, and in vitro studies, in a systemic manner, confirming ReMeDy to be a harmonized storage and visualization platform for diverse stem cell data. The relational JSON formatted database allows us to import CDE data, while employing validators for a stringent quality control.

Future aims for ReMeDy include increasing the database size to include all published iPSC and CSC research. This will be accomplished by implementing natural language processing and crowdsourcing functionalities. To automate data abstraction, we aim to use MeSH terminology and ontology-driven functionalities [8, 9]. These approaches will allow us to realize the potential of driving knowledge discovery through the use of statistical and comparative analyses of iPSC and CSC data. Crowdsourcing functionality will be implemented by expanding our iPS and CSC automated pipeline.

5. Conclusion

The ReMeDy platform allows for consolidation, harmonization, and storage of diverse stem cell CDEs, available for access in a centralized and unified manner. The platform provides the first attempt to abstract iPSC and CSC data into a single unified framework. The access to and analysis of harmonized CDEs has the potential for generation of new knowledge and advance regenerative and cancer medicine.

References

[1] Biehl JK, Russell B. Introduction to stem cell therapy. J Cardiovasc Nurs. 2009;24(2):98–105.

[2] Borziak K, Parvanova I, Finkelstein J. ReMeDy: a platform for integrating and sharing published stem cell research data with a focus on iPSC trials. Database (Oxford). 2021;2021.

[3] Parvanova I, Borziak K, Finkelstein J. A Platform for Integrating and Sharing Cancer Stem Cell Data. Annu Int Conf IEEE Eng Med Biol Soc. 2021.

[4] Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020;48(D1):D431-d9.

[5] Borziak K, Qi T, Evangelista JE, et al. Towards Intelligent Integration and Sharing of Stem Cell Research Data. Stud Health Technol Inform. 2020;272:334-7.

[6] Sakurai K, Kurtz A, Stacey G, et al. First Proposal of Minimum Information About a Cellular Assay for Regenerative Medicine. Stem Cells Transl Med. 2016;5(10):1345-61.

[7] Finkelstein J, Parvanova I, Zhang F. Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research. Stem Cells Cloning. 2020;13:1-20.

[8] Elghafari A, Finkelstein J. Introducing an Ontology-Driven Pipeline for the Identification of Common Data Elements. Stud Health Technol Inform. 2020;272:379-82.

[9] Elghafari A, Finkelstein J. Automated Identification of Common Disease-Specific Outcomes for Comparative Effectiveness Research Using ClinicalTrials.gov: Algorithm Development and Validation Study. JMIR Med Inform. 2021;9(2):e18298.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210827

Interoperability Standards for Data Sharing as a Basis to Fill in a Tailored EHR for Undiagnosed Rare Diseases

Norbert MAGGI[a,b], Ariam BOAGLIO[a], Carmelina RUGGIERO[a,c],
Roberto FANCELLU[d,e], Francesco COCCHIARA[e,f], Domenico COVIELLO[b],
Deborah CAPANNA[g] and Mauro GIACOMINI [a,c1]
[a]*Department of Informatics, Bioengineering, Robotics and Information Systems, University of Genoa, Genoa, Italy*
[b]*Laboratory of Human Genetics, IRCCS Giannina Gaslini, Genoa, Italy*
[c]*Healthropy S.r.l., Savona, Italy*
[d]*Unit of Neurology, IRCCS Ospedale Policlinico San Martino, Genoa, Italy*
[e]*Clinical Centre for Orphan Diagnosis Patients, IRCCS Ospedale Policlinico San Martino, Genoa, Italy*
[f]*Unit of Endocrinology, IRCCS Ospedale Policlinico San Martino, Genoa, Italy*
[g]*Comitato i Malati Invisibili, Genoa, Italy*

Abstract. Undiagnosed rare diseases include diseases with a well-characterised phenotype, diseases with unknown molecular causes or due to non-genetic factors, and pathological condition that cannot be named. Several initiatives have been launched for healthcare of patients with undiagnosed rare diseases. A project for development of medical records with special reference to the HL7 standards is being carried out in Genoa (Italy), taking into account regional and national regulations. The project is based on the integration of functionality related to patient diagnostics, taking into account omic sciences for disease prevention and risk assessment. Considering the evolution of standards, the use of FHIR is being considered in order to increase the elasticity of the system also in view of foreseeable adoption of this standard by the Italian healthcare system.

Keywords. Undiagnosed Rare Diseases, interoperability, standards.

1. Introduction

The Undiagnosed Rare Diseases (URDs) are conditions that describes people with a range of disorders and/or disabilities, probably caused by a genetic cause or genetic predisposition, that has not been yet identified. For these people, the lack of a definite diagnosis causes significant physiological and social consequences, with considerable diagnostic and therapeutic delays.

URDs may also include diseases with a well-characterised and described phenotype, or pathological conditions but that cannot be classified by a name, and they have an unknown molecular cause or they are due to epigenetic factors that interact with

[1] Corresponding Author, Mauro Giacomini, via all'Opera Pia, 13 – 16145 Genova, Italy; E-mail: mauro.giacomini@dibris.unige.it.

environmental factors. Several initiatives at the international level have been launched, culminating in the creation in 2014 of the NIH Undiagnosed Diseases Network, which with an interdisciplinary network of seven clinical sites has begun to make a significant impact on patients with undiagnosed rare diseases [1]. Other initiatives have been launched in Italy, such as the "Malattie senza diagnosi" (undiagnosed diseases) programme launched in 2016 by the Telethon Institute of Genetics and Medicine in Pozzuoli (Tigem) and the Clinical Centre for Orphaned Diagnosis Patients set up at the IRCSS Ospedale Policlinico San Martino in Genoa, in 2017 following an agreement with the Comitato I Malati Invisibili (Invisible Patients Association). Specifically, the latter was set up with the aim of reducing social distress, compromising quality of life and increasing co-morbidity in patients whose diagnosis is uncertain, limiting the phenomenon of diagnostic 'nomadism', and reducing the high costs to the national health system of repeatedly prescribing various types of investigation in the absence of proper coordination and critical evaluation. In order to achieve these goals, an operational protocol (clinical pathway) was developed which intends to implement IT-based communication and methodological tools for patient health management. In this respect an innovative electronic clinical record was designed, focusing on the integration of different dataset from different sources.

The requirement to use different data sources makes the use of standards essential to ensure proper interoperability [2-4]. This paper aims to describe the project for the development of a medical record with special reference to the standards that allows the use of data shared by different sources to be collected in the proposed architecture. The choice of standards will also be delimited with respect to the regulations produced by regional and national legislators [5] concerning healthcare facilities and the national/regional Health Information Infrastructure (HII).

2. Methods

The rationale behind the design of the information system dedicated to the clinical centre in Genoa is based on the integration of functionalities. In order to adequately develop the functionalities related to the clinical section and to the patient diagnostics, it will be necessary to adopt standards that will allow the complete interoperability of the medical record that will be developed with respect to the Hospital Information System (HIS), the Laboratory Information System (LIS) and Radiology Information System (RIS), also with different hospital in regional panorama.

The development of omics sciences and the increased availability of specific molecular medicine data combined with innovative DNA sequencing analysis by NGS (Next Generation Sequencing) for WES (Whole Exome Sequencing) and WGS (Whole Genome Sequencing) can improve disease prevention and risk assessment. In this respect, the use and inclusion of individual omics data and common clinical data within a tailored electronic health record (EHR) will improve disease prevention, risk assessment, treatment and diagnosis. Our approach is aimed to constitute a 'patient-centric translational EHR', relevant to the deployment of translational medicine [6], and speed up the identification of the disease. This will be based on standard classifications and terminologies such as ICD, SNOMED-CT, Orphanet, HPO and other for the development of a tool capable of identifying correlations between genotypic and phenotypic information collected and described in a relational database. This integration, together with self-learning analytical systems, may also lead to results in the development of support mechanisms for disease identification and improved accuracy of diagnosis.

Moreover, in order to achieve a better and more accurate diagnosis in a shorter time, it is envisaged that the web-based tool will allow access to specialist doctors outside the structure both for patient's clinical history consulting and to input data and observations relating to the patient. This characteristic requires data management relating to the privacy and consent of the patients, an issue that becomes even more important considering the integration of the patient's omics data [7]. In this respect, also in compliance with the provisions of the EU regulation on the processing of personal data [8], we intend to develop a privacy by design and privacy by default approach, providing from the outset the tools and the correct settings to protect personal data (Role based access control RBAC) [9]. A decision is currently being taken in coordination with the hospital authorities on whether to host the system at the hospital's data centre or to develop a cloud-based system, possibly using blockchain technology to verify the sources of and access to data [10; 11].

According to the decision of Italian legislator [5], we decide to centre our data collection system on the HL7 Clinical Document Architecture (CDA).

The use of standards will make it possible to obtain data already available in the data centres of healthcare facilities, thus filling in the data required by the physicians (Fig 1) who have commissioned this specific tailor-made EHR, achieving the goal of obtaining an early diagnosis and treatment of the patients.

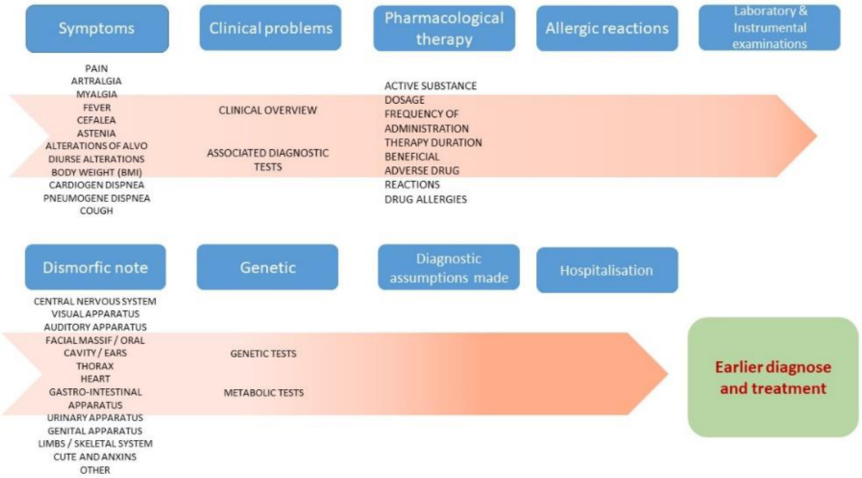


Figure 1. Patients data and processing flow for earlier diagnose EHR

3. Results

A schematic diagram of the proposed architecture is shown in figure 1, encompassing different scenarios that may arise. In a first case (a) the hospital provides the possibility of obtaining data through the use of web services, but these are not fully standardised for the production of CDA compliant documents. In this perspective a client application can be used for translation into the necessary format to ensure interoperability between the other parts of the system. In case (b), a series of views can be obtained by means of agreements with the hospital structure. An extensive discussion with the data centre managers is necessary in order to understand the logical schema with which the data is stored in the HIS. It is also necessary to interpose a standardisation client for the generation of a CDA compliant document. In the case of external specialists (c) it will be necessary to digitise the paper documentation and to generate specific meta-data describing the documents and their contents. Most of Italian regions adopted the Retrieve Locate and Update Service (RLUS) interface [2], to allow authorised entities to feed and retrieve data to and from regional HII. The presented system has been authorised to interact with HII to extract previous data and to update it.

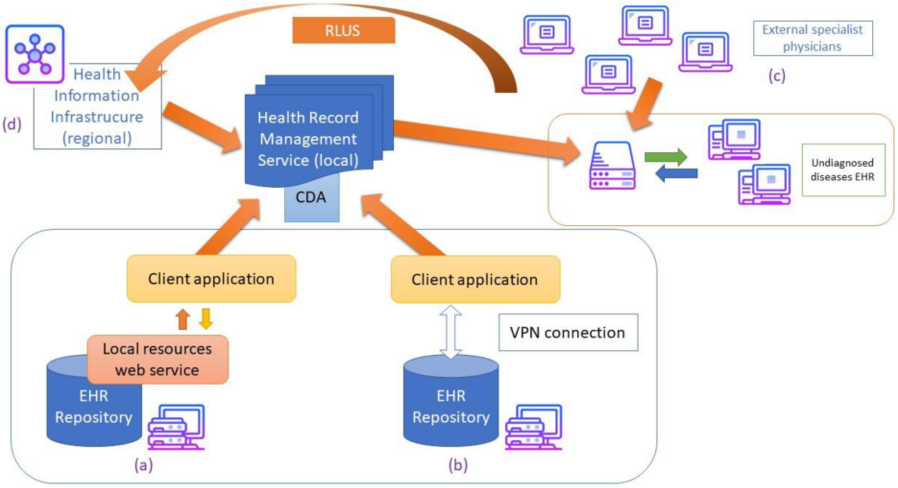

RLUS
Health Information Infrastrucure (regional)
(d)
Health Record Management Service (local)
CDA
External specialist physicians
(c)
Undiagnosed diseases EHR
Client application
Client application
Local resources web service
EHR Repository
VPN connection
EHR Repository
(a)
(b)

Figure 2. Outline of the proposed architecture

Moreover, since the patient will have interest in providing consent to access the data in the HII, the system will be able to access through the CDA the documents available at HII, if they contain analytically significant elements (d).

Considering the evolution of the standards, the use of FHIR will also be evaluated to increase the elasticity of the system, also in view of the adoption of this standard by the Italian legislator. In this respect, it is worth highlighting the introduction by HL7 of the genetic profile to the FHIR Observation resource, which will be taken into account in the development of the system [12] resulting in a future CDA and FHIR based architecture.

4. Discussion and Conclusions

Standards are an essential basis for ensuring interoperability between different information systems. The adoption of the various standards will allow the new system being developed to enable both management cooperation – different applications can interact to exchange requests and results (authentication of users with hospital credentials (LDAP)), prescriptions for drugs and analyses, discharge letters (CDA v2), booking of services through unified booking centres (HL7 v2) – and clinical cooperation. Clinical information stored in applications, even remotely managed by other healthcare professionals, can be accessed promptly at the time of need, for improved patient care.

This application could be extended later with an intensive use of Natural Language Processing (NLP) to also integrate the numerous patients' documents expressed in natural language into the information generation chain.

References

[1] Taruscio D, Baynam G, Cederroth H, Groft SC, Klee EW, Kosaki K, Lasko P, Melegh B, Riess O, Salvatore M, and Gahl WA. The Undiagnosed Diseases Network International: Five years and more!, Molecular Genetics and Metabolism. 2020; 129: 243-254.

[2] Gazzarata R, Giannini B, and Giacomini M. A SOA-Based Platform to Support Clinical Data Sharing, Journal of Healthcare Engineering. 2017;2017:1-24.

[3] Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, Sansone SA, Stoeckert CJ, Taylor CF, Taylor R, Ball CA. Data standards for Omics data: the basis of data sharing and reuse. Bioinformatics for Omics Data. 2011:31-69.

[4] D'Amore JD, McCrary LK, Denson J, Li C, Vitale CJ, Tokachichu P, Sittig DF, McCoy AB, and Wright A. Clinical data sharing improves quality measurement and patient safety. Journal of the American Medical Informatics Association 2021;28:1534-1542.

[5] DECRETO DEL PRESIDENTE DEL CONSIGLIO DEI MINISTRI 29 settembre 2015, n. 178. Regolamento in materia di fascicolo sanitario elettronico, in: Gazzetta Ufficiale Serie Generale n.263 del 11-11-2015, 2015.

[6] Shabo A, The patient-centric translational health record, Pharmacogenomics, 2013 14, 349-352.

[7] Tantoso E, Wong W-C, Tay WH, Lee J, Sinha S, Eisenhaber B, and Eisenhaber F. Hypocrisy Around Medical Patient Data: Issues of Access for Biomedical Research, Data Quality, Usefulness for the Purpose and Omics Data as Game Changer. Asian Bioethics Review. 2019;11:189-207.

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), in, 2016.

[9] Gazzarata G, Gazzarata R, and Giacomini M. A standardized SOA based solution to guarantee the secure access to HER. Conference on Enterprise Information Systems/International Conference on Project Management/Conference on Health and Social Care Information Systems and Technologies, Centeris/Projman / Hcist 2015. 2015;64: 1124-1129.

[10] Mehta S, Grant K, and Ackery A. Future of blockchain in healthcare: potential to improve the accessibility, security and interoperability of electronic health records. BMJ Health & Care Informatics. 2020:27.

[11] Sharma Y and Balamurugan B. Preserving the Privacy of Electronic Health Records using Blockchain. Procedia Computer Science. 2020;173:171-180.

[12] Alterovitz G, Heale B, Jones J, Kreda D, Lin F, Liu L, Liu X, Mandl KD, Poloway DW, Ramoni R, Wagner A, and Warner JL. FHIR Genomics: enabling standardization for precision medicine use cases. npj Genomic Medicine. 2020:5.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210828

Semantics Management for a Regional Health Information System in Italy by CTS2 and FHIR

Roberta GAZZARATA[a], Norbert MAGGI[b,c], Luca Douglas MAGNONI[a],
Maria Eugenia MONTEVERDE[a], Carmelina RUGGIERO[a,b] and
Mauro GIACOMINI[a,b,1]
[a] *Healthropy S.r.l., Savona, Italy*
[b] *Department of Informatics, Bioengineering, Robotics and Information Systems, University of Genoa, Genoa, Italy*
[c] *Laboratory of Human Genetics, IRCCS Giannina Gaslini, Genova, Italy*

Abstract. An infrastructure for the management of semantics is being developed to support the regional health information exchange in Veneto – an Italian region which has about 5 million inhabitants. Terminology plays a key role in the management of the information fluxes of the Veneto region, in which the management of electronic health record is given great attention. An architecture for the management of the semantics of laboratory reports has been set up, adopting standards by HL7. The system has been initially developed according to the common terminology service release 2 (CTS2) standard and, in order to overcome complexities of CTS2 is being revised according to the Fast Healthcare Interoperability Resources (FHIR) standard, which has been subsequently introduced. Aspects of CST2 and of FHIR have been considered in order to retain most suitable aspects of both. This integration can be regarded as most worthwhile.

Keywords. Semantic interoperability, HL7, CTS2, FHIR

1. Introduction

Over the last 20 years health care delivery has gone through significant developments, mostly relating to electronic health records and data sharing, data standards, bioinformatics and public health informatics. Health informatics technologies are normally being evaluated according to three main aspects: the ability to improve health outcomes for patients, the care quality improvement and the reduction of health costs. In USA nearly 20% of the gross domestic product is used for healthcare, and this will not be sustainable in the future, which is also applicable to the rest of the world. Digital health is being implemented in clinical practice throughout the world, and the increasing cost of digital health technologies, together with a lower extent of regulations in the related markets may result in a further expansion and acceleration of their adoption [1].

Nowadays many problems have arisen for healthcare delivery, such as infectious disease surveillance, lack of personalized care, limitations in human resources,

[1] Corresponding Author, Mauro Giacomini, via all'Opera Pia, 13 – 16145 Genova, Italy; E-mail: mauro.giacomini@dibris.unige.it..

inequitable distribution of health care. In 2018 WHO passed a resolution to develop digital health technology in order to promote equitable and universal access to health for all, and this was followed by the Global strategy on digital health 2020-2024 and by the national eHealth strategy toolkit, which was set up to help countries to integrate eHealth into their healthcare systems [2-5]. Subsequently, WHO provided recommendations on digital interventions for health system strengthening, based to address health system needs. As relates to implementation, the guideline by WHO also addresses problems for which digital health has the potential to help, such as distance and access, and shares many of the underlying challenges faced by health systems, such as poor management, infrastructural limitations and poor access to equipment [6].

Health information and communication technologies have been introduced, aiming to transform the organization of healthcare, improving quality of care and promoting access to affordable healthcare for all. At present the main modalities of digital health technologies electronic health records (EHRs), computerized provider order entry, health information exchange (HIE), Telemedicine/Telehealth, mobile-health, robots, virtual reality, wearable sensors, internet of things, artificial intelligence applications, machine learning [7].

HIE system adoption has increased worldwide in the last years, following the development and use of EHRs, which have most significant advantages with respect to paper records and whose development and use has been suggested as a key solution for the exchange of information among medical institutions and, in general, in healthcare systems [8-12]. HIE has a very high potential for health care information systems, both as relates to patient care and as relates to cost reduction for use of resources. Further research is needed to increase user participation and to develop further technology aspects [13]. Data sharing is a key building block for effective healthcare delivery. The main interacting systems that manage patient's data and could provide data for HIEs are EHRs, which store clinical information, such as patient's medical history, diagnoses, medications, laboratory results, which store and manage clinical laboratory data, and picture archiving and communication systems, which store and manage medical images.

Interoperability in eHealth has been addressed by the European Union (EU), which has set up the Refined eHealth European Interoperability Framework (EIF), which considers many different aspects of interoperability [14; 15].

An infrastructure for semantic interoperability is being developed to support the regional HIE in Veneto – an Italian region which has about 5 million inhabitants. This infrastructure aggregates data according to data semantics. The management of semantics is one of the key aspects of HIEs, because in medical practice terminologies used in different departments, laboratories and institutes are usually diverse and very different from standardized vocabularies, while standardized terminologies, universally recognized for each specific application domain, should be adopted [16-18].

2. Methods

The Logical Observation Identifiers Names and Codes (LOINC) [19] vocabulary has been used, in order to represent concepts and relations among concepts which are defined in different local and standardized terminologies. LOINC is frequently updated, in order to maintain technologies and their relations up-to-date and coherent over time [20]. Concepts and terminologies relating to several laboratory tests in the Veneto Region have been encoded by LOINC. The results have been stored in a database and can be downloaded by a table containing all laboratory tests and the related LOINC entities. The table has been uploaded in LISs, therefore LOINC codes have been used in the Clinical Documents Architecture (CDA) laboratory reports.

According to the recommendations by the Italian Health Ministry, standards by HL7 have been adopted. The CTS2 standard provides specifications to develop interfaces to manage, search and access terminology contents. CTS2 has been set up within the HL7 and Object Management Group initiative by the Healthcare Service Specification Project (HSSP) [20]. HSSP aims to define industry standards based on SOAs to achieve interoperability among applications that belong to independent socio-health system organizations [17; 21-23]. CTS2 defines elements called terminology resources and sets of operations, called functional profiles, which could be performed on them [24; 25].

In order to overcome the complexities of CTS2, HL7 has subsequently introduced the Fast Healthcare Interoperability Resources (FHIR), a standard aiming to improve healthcare information exchange using building blocks – called resources - which define common concepts, that is small units of data, such as observation, condition, device, patient. Resources increase the reusability of health information and are intended to cover typical use cases [26-28]. FHIR is increasingly adopted by technology companies and might see a faster adoption than other standards [27].

A terminology service has been developed for the Veneto, initially according to the CTS2 reference model. Subsequently the service has been integrated into a FHIR based system for terminology management, in order to improve speed and information reusability. Terminology plays a key role in the management of the information fluxes of the Veneto – in which the management of EHRs is given great attention. The adoption of a FHIR interface in the developed terminology system was due to the fact that Veneto Region adopted FHIR as its main sematic signifier for its Health Information Infrastructure (HII). Moreover, FHIR interface also improved the performances of the presented terminology system.

The CTS2 terminology resources that have been used are CodeSystem, CodeSystemVersion, EntityDescription, Map, MapVersion, MapEntry. For these elements the functional profiles Maintenance, Read, Query, History have been considered. The implementation profile that has been chosen is the Simple Object Access Protocol (SOAP), and the system has been hosted in Microsoft Windows Azure [18].

The system is being revised considering aspects of CTS2 and of FHIR, in order to retain the most suitable aspects of both to improve speed and reusability of information.

3. Results and Discussions

The architecture for the management of the semantics of laboratory reports is shown in figure 1. Its main components are the Health Terminology Service (HTS), the client web application for the management of the information in the HTS, the Laboratory Information Systems (LISs) of the regional departments and regional HIE. The main component of the architecture is the HTS, which consists of a relational database in which all information relating to terminology resources is stored, and of a set of web services compliant with the CTS2 standard. The relational database is hosted in Microsoft SQL Azure. A set of web services provides access to the database by a CTS2 interface, consisting of a set of Windows Communication foundation (WCF) services hosted in Microsoft Azure [18]. Each terminology resource has one service for each functional profile, therefore the resulting HTS has 24 WCF services. Therefore, the FHIR model has been adopted, which allows to replace the 24 WCF services with one FHIR resource. This significantly improves the speed of the system.

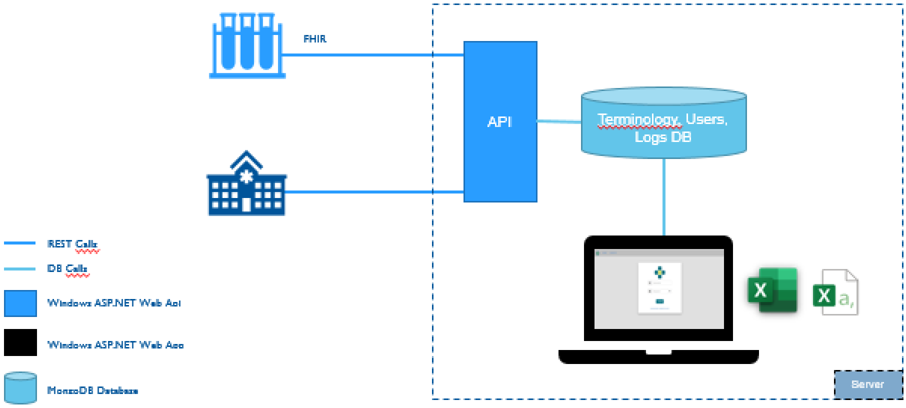


Figure 1. Architecture scheme

Other aspects have also been considered. The central database has CTS2 objects which resemble the FHIR ones, such as, the state of codes, which can be active or not. Moreover, a service which transmits FHIR messages has been added to the interface.

In conclusion, the integration of aspects of CTS2 and of FHIR can be regarded as most worthwhile. Further developments along these lines are being considered.

4. Conclusions

The new HTS architecture based on FHIR message is still under test in work environment in Veneto Region, but preliminary results seem to be very promising, both as regards the speed performance of the system and for the capability of the system to maintain history of the terminology data sets, even at the concept/term level. This is probably due to the correct mixed used of the CTS2 features and of the FHIR specificities.

The correct use of a terminology system will significantly improve the use of HII that could be a fundamental tool to assure patients continuity of care and strengthen the delivery of territorial healthcare services.

References

[1] Rivas H, Creating a Case for Digital Health, in: Digital Health, 2018, pp. 1-13.
[2] Dhingra D and Dabas A. Global Strategy on Digital Health. Indian Pediatrics. 2020; 57: 356-358.
[3] World Health Organization, mHealth: use of appropriate digital technologies for public health: report by the Director-General., in, World Health Organization: Genève, Switzerland, 2017.
[4] World Health Organization, Global strategy on digital health 2020-2025, World Health Organization: Genève, Switzerland, 2021.
[5] World Health Organization, National eHealth strategy toolkit, in, World Health Organization: Genève, Switzerland, 2012.
[6] World Health Organization, WHO guideline: recommendations on digital interventions for health system strengthening: evidence and recommendations, in World Health Organization: Genève, Switzerland, 2019.

[7] Stein AT, Ben ÂJ, Pachito DV, Cazella SC, van Dongen JM, and Bosmans JE. Digital Health Technology Implementation: Is It Effective in a Healthy Healthcare Perspective?, in: Integrating the Organization of Health Services, Worker Wellbeing and Quality of Care. 2020:197-220.

[8] Warren LR, Clarke J, Arora S, and Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. BMJ Open. 2019:9.

[9] Jones SS, Rudin RS, Perry T, and Shekelle PG. Health Information Technology: An Updated Systematic Review With a Focus on Meaningful Use. Annals of Internal Medicine. 2014:160.

[10] Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, Searcy T, and Jha AK. Electronic Health Record Adoption In US Hospitals: Progress Continues But Challenges Persist. Health Affairs. 2015;34:2174-2180.

[11] Ji H, Kim S, Yi S, Hwang H, Kim J-W, and Yoo S. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. Journal of Biomedical Informatics. 2020;107.

[12] Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, and Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearbook of Medical Informatics. 2017;26: 28-37.

[13] Sadoughi F, Nasiri S, and Ahmadi H. The impact of health information exchange on healthcare quality and cost-effectiveness: A systematic literature review. Computer methods and programs in biomedicine. 2018;161:209-232.

[14] Spanakis EG, Sfakianakis S, Bonomi S, Ciccotelli C, Magalini S, and Sakkalis V. Emerging and Established Trends to Support Secure Health Information Exchange. Frontiers in Digital Health. 2021;3.

[15] European Commission Directorate-General for Informatics, New European interoperability framework - Promoting seamless services and data flows for European public administrations, in, 2017.

[16] Canepa S, Roggerone S, Pupella V, Gazzarata R, and Giacomini M. A Semantically Enriched Architecture for an Italian Laboratory Terminology System, in: XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013. 2014:1314-1317.

[17] Gazzarata R and Giacomini M. A Standardized SOA for Clinical Data Sharing to Support Acute Care, Telemedicine and Clinical Trials. European Journal for Biomedical Informatics. 2016:12.

[18] Gazzarata R, Monteverde M, Vio E, Saccavini C, Gubian L, Borgo I, and Giacomini M. A Terminology Service Compliant to CTS2 to Manage Semantics within the Regional HIE. European Journal for Biomedical Informatics. 2017:13;43-50.

[19] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, and Maloney P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. 2003:49;624-633.

[20] Kawamoto K, Honey A, and Rubin K. The HL7-OMG Healthcare Services Specification Project: motivation, methodology, and deliverables for enabling a semantically interoperable service-oriented architecture for healthcare. J Am Med Inform Assoc. 2009;16:874-881.

[21] Gazzarata R, Vergari F, Cinotti TS, and Giacomini M. A standardized SOA for clinical data interchange in a cardiac telemonitoring environment. IEEE J Biomed Health Inform. 2014;18: 1764-1774.

[22] Gazzarata G, Gazzarata R, and Giacomini M. A standardized SOA based solution to guarantee the secure access to EHR, Conference on Enterprise Information Systems/International Conference on Project Management/Conference on Health and Social Care Information Systems and Technologies, Centeris/Projman / Hcist 2015. 2015;64:1124-1129.

[23] Gazzarata R, Giannini B, and Giacomini M. A SOA-Based Platform to Support Clinical Data Sharing. Journal of Healthcare Engineering. 2017;2017:1-24.

[24] Hamm R, Estelrich A, Canu N, Oemig F, and Nachimuthu S. HL7 Common Terminology Services, Release 2: Service Functional Model Specification, Normative Release, in, Health Level Seven International: Ann Arbor, MI, USA, 2015.

[25] OMG, Common Terminology Services 2, Version 1.2, in, OMG: Needham, MA, USA,, 2015.

[26] HL7.org, FHIR overview, in.

[27] Lehne M, Luijten S, Vom Felde Genannt Imbusch P, and Thun S. The Use of FHIR in Digital Health - A Review of the Scientific Literature. Stud Health Technol Inform. 2019;267:52-58.

[28] Lee AR, Kim IK, and Lee E. Developing a Transnational Health Record Framework with Level-Specific Interoperability Guidelines Based on a Related Literature Review. Healthcare. 2021;9.

doi:10.3233/SHTI210829

What Metadata? Defining Different Types of Digital Assets as Application Targets of Metadata in Clinical Research Informatics

Matthias LÖBE[a,1]
[a] *Institute for Medical Informatics (IMISE), University of Leipzig, Germany*

Abstract. The term 'metadata' is mentioned in every one of the FAIR principles. Metadata is without question important for findability, accessibility, and reusability, but essential for interoperability. Standardized schemas have been developed by various stakeholders for decades, but too rarely come to practical use. The reason for this is that the application domain is not clearly understood. In many bio-medical research projects, the need for metadata is recognized at some point, but there is not only a lack of overview of existing standards, but also a lack of correct assessment of what individual metadata schemas were actually made for. This paper differentiates different application scenarios for metadata in clinical research.

Keywords. Metadata, Controlled Vocabularies

1. Introduction and Method

Metadata has a rather mystical meaning for many clinical researchers. The term itself does not have an undisputed definition, often scientists talk of "data about data", which on the one hand is rather general, on the other hand excludes objects that are not data. Computer scientists value metadata as necessary artifacts to exchange data between information systems without loss of information. Especially in the field of clinical research, data collection is carried out with great human, financial and regulatory cost, and there is a growing awareness that clinical studies should be registered prospectively and that results should be published even if they fail. Furthermore, there are increasing voices calling for the leakage of primary data, including individual patient data in de-identified form, to appropriate researchers.

Many research groups therefore store data sets and documents that form the basis for publications in central research data repositories, in which access can be granularly regulated according to protection needs. However, the pure instance data are only valuable for secondary analyses if the collected medical concepts behind the variables and the structure of the conducted research can be interpreted. A variety of different standards, metadata vocabularies, and medical terminologies are available for this purpose - both applicable to generic digital assets and specific to particular subfields such as clinical trials or health care data. However, if one looks at the practical use of

[1] Corresponding Author, Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de.

standardized metadata schemas, for example among the 2,700 research data repositories listed under the meta registry re3data [1], only a fraction uses generic metadata schemas such as Dublin Core or DataCite. Subject-specific vocabularies are even rarer by orders of magnitude. The hypothesis of this work is that many researchers do not sufficiently consider what type of digital assets they want to describe in the first place and therefore do not really use appropriate standards, which then have to be modified and extended, ultimately limiting interoperability. Using an expert-based approach, different types of assets were identified and organized.

2. Results and Discussion

Three main groups and nine subgroups of digital assets in clinical studies can be distinguished, for which very different metadata standards are relevant:

1. Structural description of data objects (structured data or documents)
    1.1. Design of the experiment (arms, cohort definition, endpoints, study sites)
    1.2. Timing of the experiment (phases, collection events)
    1.3. Structure of data collection (data models, forms, instruments, item groups, data elements, code lists)
2. Administrative description primarily for research data management
    2.1. Projects, agents, and stakeholders
    2.2. Data sets (databases) and data distributions (snapshots)
    2.3. Information systems (portals, repositories) and the catalogs they contain
3. Annotation for data usage
    3.1. Provenance (data origin, transformations, measuring methods)
    3.2. Data quality (validation and curation)
    3.3. Availability (restrictions on reuse: legal basis, patient consent)

For all these groups, metadata schemas can be found that ensure a widely accepted semantic foundation through internationally agreed standards and medical terminologies. Several candidates exist for each group; however, explaining and classifying them is beyond the scope of this paper and is the goal of future work. However, it is important to choose a suitable standard that fits the corresponding group. Otherwise, there will quickly be a need for project-specific modifications and extensions that will hamper true interoperability. Better than overambitious in-house developments is the use of coordinated vocabularies as stated in FAIR Principle R1.3: (Meta)data meet domain-relevant community standards [2], in order to develop best practices of the application of precise metadata elements in the medium term and to keep the effort for submitters low as well as for consumers.

Acknowledgments: This work was supported by the German Research (DFG grants WI 1605/10-2) and the European Commission (H2020 grant 824666).

References

[1] re3data.org - Registry of Research Data Repositories. https://doi.org/10.17616/R3D

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, et.al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18.

This page intentionally left blank

Section V

Paradigms to Share Health Research Data & Various Health Informatics Studies

This page intentionally left blank

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210831

Making EHRs Reusable: A Common Framework of Data Operations

Miguel PEDRERA[a,b,1], Noelia GARCIA[a], Paula RUBIO[a], Juan Luis CRUZ[a], José Luis BERNAL[a] and Pablo SERRANO[a]
[a]*Hospital Universitario 12 de Octubre, Madrid, Spain*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain*

Abstract. Reuse of EHRs requires data extraction and transformation processes are based on homogeneous and formalized operations in order to make them understandable, reproducible and auditable. This work aims to define a common framework of data operations for obtaining EHR-derived datasets for secondary use. Thus, 21 operations were identified from different data-driven projects of a 1,300-beds tertiary Hospital. Then, ISO 13606 standard was used to formalize them. This work is the starting point to homogenize ETL processes for the reuse of EHRs, applicable to any condition and organization. In future studies, defined data operations will be implemented and validated in projects of different purposes.

Keywords. Electronic Health Records, FAIR, Data reusability, Real World Data, Semantics, Standards, ISO 13606, i2b2, OMOP, ISARIC, COVID-19.

1. Introduction

Electronic Health Record (EHR) is defined as the repository of health data that is generated throughout a patient's lifetime. Its primary use is to enable continuous, efficient and quality healthcare [1]. Additionally, there are other uses of EHR, known as secondary uses, including activities such as clinical research or public health [2]. These further uses are only possible if we produce reusable EHR data, which is one of the principles established by FAIR [3].

Reusability is determined by how we manage the semantics of concepts and (meta)data in information systems. A first step is provided by Detailed Clinical Models (DCM), which allow implementing mechanisms for obtaining EHR-derived datasets for secondary use [4, 5]. However, it is essential that the extraction, transformation and loading processes (ETLs) are based on homogeneous and formalized operations, in order to make them understandable, reproducible and auditable [6].

Thus, this work aims to define a common framework of data operations on EHRs, necessary for them to be adequately reusable for secondary purposes.

[1] Corresponding Author, Miguel Pedrera Jiménez, Hospital Universitario 12 de Octubre, Av. de Córdoba s/n, 28041 Madrid Spain; E-mail: miguel.pedrera@salud.madrid.org.

2. Methods

This work was carried out at Hospital Universitario 12 de Octubre (H12O) in Madrid (Spain), as part of its research line on the effective reuse of EHRs [4, 7, 8].

2.1. Detailed Clinical Models

In this study, DCM were used as the basis for the design and formalization of data operations. This paradigm proposes a dual model composed of a reference model and an archetype model [9]. Thus, ISO 13606 standard [10], previously adopted by H12O and Spanish Ministry of Health, was selected for this purpose.

The reference model of this standard defines the components for building an interoperable EHR: Folder, Composition, Section, Entry, Cluster and Element. It also establishes the types of data permitted, which allows limiting the valid data types for each operation. In the present work, it was necessary to use the following subset:

- **Coded Value (CV)**: for concepts whose result is a set of possible coded values, e.g., SARS-COV-2 test, which may be positive, negative or inconclusive;
- **Physical Quantity (PQ)**: for concepts whose outcome is a numerical value with unit of measurement, e.g., oxygen flow rate measured in liters per minute;
- **Integer**: for concepts whose result is an integer value, e.g., Glasgow Comma Scale score; and,
- **Date Time**: for concepts whose value is a time point, e.g., date of symptom onset.

Likewise, the archetype model allows the information models to be formalized and linked with terminologies at two levels: the semantic binding, for specifying the meaning of its components, and the value binding, to define the set of values of a CV element.

A reusable EHR must be supported by appropriate modeling and standardization practices. Therefore, it is necessary to use common reference models and standard terminologies, such as SNOMED CT [11] and LOINC [12], that do not contain miscellaneous, grouped, calculated or inferred concepts. Thus, the execution of formalized data operations on standardized EHR extracts (structure and content) allows ETL process to be applicable regardless of the condition and organization.

2.2. Secondary use data models

Secondary use models allow data to be represented and persisted for other uses in addition to healthcare. Consequently, they are less demanding than primary use models in terms of metadata about the registration process or access permissions. We can distinguish two types of secondary use models:

- **Clinical data repositories.** These models centralize data from different sources within a common structure and content. They have not been modeled for a single purpose, but as a data warehouse for multiple secondary uses, e.g., i2b2 (tranSMART Foundation) [13], used in TriNetX Platform (federated network for clinical trials) [14], and OMOP CDM (OHDSI) [15], used in EHDEN Consortium (federated network for observational research) [16].
- **Electronic Data Capture systems (EDC).** These models collect data as it is expected to be analyzed. They are designed according to specific use cases, e.g.,

ISARIC Case Report Form (CRF) for COVID-19 [17] and STOP-CORONAVIRUS EDC [18].

In order to define the set of common data operations, the different models that have been used for different data-driven projects in H12O were analyzed, considering both typologies. Table 1 shows the list of specifications reviewed.

Table 1. Data-driven projects analyzed for identification of data operations.

ID	Data-driven project	Data model typology	Purpose
1	TriNetX Platform	i2b2 repository	Clinical Trials and analytics
2	EHDEN Consortium	OMOP repository	Observational studies
3	ISARIC Consortium	Specific EDC	Case reports and analytics
4	STOP-CORONAVIRUS	Specific EDC	Observational studies

2.3. Identification and formalization of data operations

Once the data models of the different projects have been analyzed, data operations were identified and then classified according to several categories. These high-level operations, parents of the fully defined operations (FDO), were as follows:

- **Selection (S).** Operations to select and extract the required data under the restrictions of the secondary use model. Two subtypes were defined:
  - **Selection with reference (S.1),** e.g., selection of "Oxygen saturations" less than 96%.
  - **Selection without reference (S.2),** e.g., selection of the "Oxygen saturation" with the lowest value.
- **Transformation (T).** Operations to transform the data to the format of the secondary use model. Two subtypes were defined:
  - **Transformation maintaining meaning (T.1),** e.g., changing the measurement unit of a concept "C-Reactive Protein" from mg/dL to mg/L.
  - **Transformation altering meaning (T.2),** e.g., calculating a "BMI" concept from "Weight" and "Height".

Operations were formalized by specifying the valid data types and the cardinality of the argument, input and output of them. For this purpose, the data types specified in the ISO 13606 reference model were employed.

3. Results

3.1. Identification of data operations

The first result obtained was the set of data operations, classified according to the categories defined in the methodology section. Table 2 shows this specification, indicating, for each FDO, an example and the projects that required them.

Table 2. Identified data operations for EHRs reuse.

ID	Operation	Example	Project
S	Selection	-	-
S.1	Selection with reference	-	-
S.1.1	Selection of data related to *concept*	Data related to COVID-19 test results	All

S.1.2	Selection of data previous to *date*	Pre-hospitalization medication	All
S.1.3	Selection of data after *date*	Medication during hospitalization	All
S.1.4	Selection of data higher than *value*	Temperatures higher than 37 ºC	3, 4
S.1.5	Selection of data less than *value*	Oxygen saturations less than 96%	3, 4
S.1.6	Selection of data equal to *value*	COVID-19 test results equal to 'Positive'	3, 4
S.2	Selection without reference	-	-
S.2.1	Selection of most recent datum	Last COVID-19 test result	3, 4
S.2.2	Selection of oldest datum	First Oxygen saturation on admission	3, 4
S.2.3	Selection of datum with higher value	Higher Temperature	3, 4
S.2.4	Selection of datum with lower value	Lower Oxygen saturation	3, 4
T	Transformation	-	-
T.1	Transformation maintaining meaning	-	-
T.1.1	Change of unit of measure	C-Reactive Protein from mg/dL to mg/L	All
T.1.2	Change of coding system,	Cough from local code to SNOMED CT	All
T.2	Transformation altering meaning	-	-
T.2.1	Mathematical operation	BMI from Weight and Height	3, 4
T.2.2	Semantic inference	Fever from Temperature	3, 4
T.2.3	Event count	Number of previous hospitalizations	3, 4

3.2. Formalization of data operations

The second result was the formalized set of FDO. To this end, data types for argument, input and output of operations were specified according to ISO 13606, as well as the cardinality (arguments have unique cardinality). Table 3 shows this specification.

Table 3. Formalized data operations for EHRs reuse.

Operation ID	Argument Datatype	Input Data type	Output Data type	Input Card.	Output Card.
S.1.1	CV	All	Same than Input	1..N	1..N
S.1.2	DATETIME	All	Same than Input	1..N	1..N
S.1.3	DATETIME	All	Same than Input	1..N	1..N
S.1.4	PQ, INTEGER	PQ, INTEGER	Same than Input	1..N	1..N
S.1.5	PQ, INTEGER	PQ, INTEGER	Same than Input	1..N	1..N
S.1.6	CV, PQ, INTEGER	CV, PQ, INTEGER	Same than Input	1..N	1..N
S.2.1	-	All	Same than Input	1..N	1..1
S.2.2	-	All	Same than Input	1..N	1..1
S.2.3	-	PQ, INTEGER	Same than Input	1..N	1..1
S.2.4	-	PQ, INTEGER	Same than Input	1..N	1..1
T.1.1	-	PQ	PQ	1..1	1..1
T.1.2	-	CV	CV	1..1	1..1
T.2.1	-	PQ, INTEGER	Same than Input	1..N	1..1
T.2.2	-	All	All	1..N	1..1
T.2.3	-	All	INTEGER	1..N	1..1

4. Conclusions

In this study, a common framework of data operations was theoretically defined for obtaining secondary use models from EHRs. For this purpose, four data-driven projects in which H12O participates were studied (Table 1).

Thus, 21 operations were identified, 15 of which were FDO (Table 2). Data models related to standardized repositories did not involve complex operations. However, specific data models for COVID-19 research required selections with complex criteria and meaning-altering transformations. The set of FDO was formalized (data types and cardinality) using ISO 13606 standard reference model (Table 3). This allows implementing homogeneous ETL processes based on common criteria and identifying processes with inconsistent operations (e.g., a 'unit change' operation on a CV variable). Moreover, these operations can be adapted in accordance to data sources and secondary use models, being applicable to other organizations and health conditions.

In future studies, data operations will be implemented with programming languages such as R, and validated in COVID-19 projects and studies of other clinical conditions.

Acknowledgment

This work has been supported by Research Projects PI18/00981 and PI18/01047 funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

References

[1] Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inform 2008;77:291–304. doi:10.1016/j.ijmedinf.2007.09.001.

[2] Safran C, Bloomrosen M, Hammond E, et al. Toward a National Framework for the Secondary Use of Health. Jounal Am Med Informatics Assoc 2007;14:1–9. doi:10.1197/jamia.M2273.Introduction.

[3] FAIR Principle R1.3. https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/. Accessed July 30, 2021.

[4] Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform. 2021;115:103697. doi:10.1016/j.jbi.2021.103697.

[5] Lim Choi Keung S, Zhao L, Rossiter J, et al. Detailed clinical modelling approach to data extraction from heterogeneous data sources for clinical research. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci 2014;2014:55–9. doi:10.1016/j.ic.2014.12.007.

[6] Kohane IS, Aronow BJ, Avillach P, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021;23(3):e22219. Published 2021 Mar 2. doi:10.2196/22219.

[7] Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a Tertiary Hospital During COVID-19 Pandemic: A Multi-Purpose Approach Based on Standards. Stud Health Technol Inform. 2021;281:28-32. doi:10.3233/SHTI210114.

[8] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. Stud Health Technol Inform. 2020;270:78-82. doi:10.3233/SHTI200126.

[9] Beale T. Archetypes: Constraint-based Domain Models for Future- proof Information Systems. OOPSLA 2002 Work Behav Semant 2001;:1–69. doi:10.1.1.147.8835.

[10] ISO 13606 standard. https://www.iso.org/standard/67868.html. Accessed July 30, 2021.

[11] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279-290.

[12] McDonald CJ, Huff SM, Suico JG, *et al.* LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin Chem* 2003;49:624–33. doi:10.1373/49.4.624.

[13] Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-130. doi:10.1136/jamia.2009.000893.

[14] TriNetX Platform. https://trinetx.com/. Accessed July 30, 2021.

[15] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-578.

[16] EHDEN Consortium. https://www.ehden.eu/. Accessed July 30, 2021.

[17] ISARIC-WHO CRF for COVID-19. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/. Accessed July 30, 2021.

[18] STOP-CORONAVIRUS. https://imas12.es/blog/stop-coronavirus-nuevo-proyecto-clinico-llevado-a-cabo-en-el-instituto-i12-para-ofrecer-respuestas-integrales-a-la-covid-19/. Accessed July 30, 2021.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210832

Utilising the FOXS Stack for FAIR Architected Data Access

John MEREDITH[a,1], Nik WHITEHEAD[b] and Michael DACEY[b]
[a] *Wales Institute for Digital Information, Cardiff UK*
[b] *University of Wales Trinity Saint David, Swansea UK*

Abstract. A FOXS stack assembles HL7 FHIR, openEHR, IHE XDS and SNOMED CT as an operational clinical data platform to build digital systems. This paper analyses its applicability for FAIR-enabled medical research based on a summary of key principles. It highlights the benefit of the blended approach to operational technology stacks for health systems, and a need for industry standard technologies to enable greater semantic coherence for primary/secondary data use.

Keywords. interoperability, HL7 FHIR, openEHR, IHE XDS, SNOMED CT, FAIR.

1. Introduction

There is now a paradigm shift in health data accessibility to the requirement for fully structured, semantically coherent data available via open APIs. The open architecture approach is key to the meaningful use of data for operational clinical tools as well as providing a foundation for secondary use for research purposes. The FAIR Principles to ensure data are *Findable*, *Accessible*, *Interoperable* and *Reusable* [1] have been adopted by organisations such as Health Data Research UK, who consider embracing open standards as a necessity [2]. However there has been little requirement for FAIR to be adopted by the UK NHS outside of research.

In this paper, we present an approach to supporting FAIR principles for health data stemming from the open clinical data interoperability platform; the FOXS stack. This employs four commonly used technology standards within the domain of digital health; HL7 FHIR, openEHR, IHE XDS and SNOMED CT. These are specifications and technologies to persist clinical data, bound by standardised terminologies, represented by syntax and metadata harmonised messaging and document structures. We present and discuss the high-level summary of each FOXS component and assess compatibility with FAIR principles.

1.1. The FOXS Stack

The FOXS stack is assembled from the following components:

- FHIR: Fast Healthcare Interoperability Resources (FHIR) from HL7 are regarded as the emerging standard for technical and syntactic interoperability [3],

[1] Corresponding Author, John Meredith, Wales Institute for Digital Information, 21 Cowbridge Road East, Cardiff CF11 9AD, UK; E-mail: john.meredith@wales.nhs.uk.

- openEHR is a specification [4] that describes clinical models and the rules which govern them. It is constrained by a reference model to support the longitudinal record, and acts as the core data persistence layer of FOXS,
- XDS: Cross-Enterprise Document Sharing (XDS.b) from IHE is a standards-based specification [5] to support the sharing of documents and images between health organisations.
- SNOMED CT is a hierarchal clinical vocabulary for use with digital tooling and patient records used widely across the world [6], mappable to other code systems and bound within data structures used by the above.

Both FHIR and openEHR specifications are freely available under open-source licenses. However, it is essential that standards have been adopted for use by the healthcare providers. While it is possible to create standardised structures for documents in both openEHR and FHIR, neither has demonstrated implementation at scale that rivals the more mature standard of IHE XDS.b for interoperability across the health sector [7].

2. Methods

The applicability of utilising some FOXS components for FAIR research has been established. NIH has issued a RFI [8] on clinical research utilising FHIR indicating a desire to test its efficacy. Recent progress also includes work groups aiming to harmonise FHIR with the BRIDG reference model [9]. Compliance to FAIR principals has also been established for openEHR [10]. This notion is extended here to the generalised FOXS stack based upon previously published implementation considerations [11].

Certain principles were established for the *definition* of metadata within FOXS stack to delineate between provenance and knowledge-based data. Provenance assumes an inherent ownership and position within the clinical pathway and can be found in FHIR, openEHR and XDS.b specifications. Knowledge metadata is attributed to ontological aspects of openEHR archetypes and FHIR based resources, to differing levels of detail.

For example, the clinical model for blood pressure exists as an openEHR archetype but described within the contents of a FHIR observation resource. While openEHR classes are analogous to certain FHIR resources, openEHR offers richer published definitions to support knowledge metadata for research. Any given model may contain references to external SNOMED CT codes (e.g. a ‘record artifact’ based document type) as well as contained within the data itself. In this scenario, we consider openEHR, FHIR and SNOMED CT with capabilities to describe clinical content, as well as metadata elements. IHE XDS.b represents a standardised container used in healthcare for additional clinical content such as documents or images (e.g. using the DICOM standard).

3. Results

Each FAIR principle is summarised with compliance to FOXS components as a whole.

3.1. Findability

Unique and persistent identifiers are the basis for *findability* to ensure computability. FHIR, openEHR and XDS rely upon location-based URLs for resources, compositions and documents respectively. OpenEHR facilitates metadata contained within a composition structure to be located via this identifier. FHIR presents optional metadata elements within individual resources. XDS utilises metadata structures as part of an *affinity domain* to locate records which may consist of a variety of different data standards for content (e.g. FHIR).

In terms of metadata richness, FHIR resources contain mandatory backbone elements and referenced Resources such as 'Patient' and 'Encounter' as well as optional metadata elements. OpenEHR attributes metadata at multiple levels to consider the provenance of the data as well as the models themselves, akin to FHIR but in more granular detail owing to the archetype modelling approach. The combination of FOXS components facilitates data descriptors identified by class of archetype or resource, document type, care context or clinical term. Computability is made possible with content due to the nature of archetypes, resources and terms being unique identifiers to data. (Meta)data may be stored separately in model repositories such as CKM [12].

3.2. Accessibility

FOXS stack components rely on W3C standard web-based protocols such as SOAP XML (XDS) and REST (FHIR, openEHR). FHIR offers a framework for search capabilities and within openEHR architecture, all data items attributed to a specific archetype may be queried independently with the archetype query language (AQL, [8]). This presents as an increasingly granular capability as query use cases move from the higher-level interoperability space (FHIR, XDS.b) to the data persistence layer (openEHR), with both perspectives augmented by SNOMED CT.

Role base access and authentication is supported with rich metamodels. Rules on persistence and retention may be subject to domain specific policy such as Caldicott data sharing principles in the UK[13]. Maintaining these structures to support *Findability* also facilitates *Accessibility* when record management policy has a consideration for the destruction of data when no longer needed [14], as this allows metadata to persist longer.

3.3. Interoperability

Utilising archetypes as the base model for FOXS facilitates semantic coherence and a common representation of knowledge. This may be demonstrated through a variety of widely used, machine-readable formats (e.g. XML, JSON, RDF). OpenEHR provides an internal terminology and data may be enhanced with the use of SNOMED CT to act as a vocabulary. In addition, FHIR has successfully been utilised to represent SNOMED CT reference data sets[15] which can in turn be embedded within openEHR models. The addition of XDS.b completes the interoperable capability for document bound data.

3.4. Reusability

The final principle concerns how (meta)data are reused and their applicability to clinical use cases. This requires implicit understanding of the use and misuse of data supported by robust and detailed metadata. A FOXS-based platform is able to make use of conformance archetypes that map directly to the interoperability layer (e.g. matching an XDS.b profile). Data usage may be encapsulated within a consent focussed archetype that describes specific scenarios such as research. Provenance-based (meta)data enables operational and secondary use filtering (e.g. only include observations recorded recently or in specific clinical contexts). Minimal data standards may be reflected as cardinality within an archetype/resource, or at the conformance API layer (i.e. a FHIR profile).

4. Discussion

FOXS implementations view SNOMED CT as a component of the wider stack, rather than it being the ontology itself. Wider structures provided by the persistent archetype model provide each data point. The increased ontological context for the uniquely identifiable nature of clinical data via a cumulatively standards based approach enables the data to become FAIR [16]. While it is feasible to include multiple terminologies or classifications within openEHR archetypes and FHIR resources, SNOMED CT becomes an invaluable tool due to its capability to map to external terminologies such as ICD-10. This is essential for enabling FAIR enabled research [17].

While the interoperable aspects of openEHR, through the native API service layer, comply with FAIR principles, it could be argued that it lacks formal standardisation in terms of industry usage. This is because the openEHR specification enables standards for clinical data models to be created that reflect use cases at the data persistence layer. The commonality of shared archetypes does not imply that the assembly of said models will be standard across all implementations. The resulting templates and APIs will reflect the decisions taken by local implementers to support the specific use case at hand. This also applies to FHIR (and to some extent XDS and SNOMED CT) that all are subject to various levels of localised standardisation to reflect the heterogenous practice across organisation or geographical boundaries.

Efforts have been made to align FAIR principles to FHIR [18], however progress has been hampered due to existing gaps between health and research standards for basic model elements such as demographics [19]. FHIR is regarded as the messaging standard for health, and is seeing increased use in the UK through initiatives such as the UK Core [20]. This advocates a desire for compliance to accessibility standards consumed by operational systems, through a common syntax. Where these interoperable APIs share common data structures, they may be used by multiple systems or actors, but also customised to support local implementation requirements. This necessitates a degree of transformation between systems, producing non-standardisation that would seem to oppose some FAIR principles such as accessibility. Nevertheless, the ability to rely on a common, persistent baseline model in openEHR, supported by SNOMED CT terms as a common vocabulary enables a FAIR-enabled technology stack to reside within operational digital systems and not exist as a purely research-based endeavour. Additionally, openEHR acts as a proxy for common data elements, essential for FAIR data sharing [19]. By abstracting this model away from the requirements of messaging, openEHR compliments FHIR, providing the flexibility demanded by implementors while maintaining the semantic coherence that benefits FAIR.

We suggest that by utilising FHIR as the interoperability gateway, we begin to standardise FAIR principles alongside established industry practice. Relying on XDS profiles to support document and image archive-based paradigms, supported by a robust metadata model supports this view.

5. Summary

The assembly of FOXS components represents an act of domain-based convergence. It juxtaposes the detailed curation of clinical models for data persistence with interoperability, utilising standard syntax and protocols. While this generalises at the enterprise scale, the assembly attempts to enable FAIR-ness by way of facilitating data access through one or more routes within a FOXS platform. Future research will seek to develop a specification for a FAIR-enabled FOXS stack and assess how this aligns to currently available FAIR maturity models [21] to support clinical data research.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 31 [Internet]. 2016 Mar 15 [cited 2021 Jul 21];3(1):1–9. Available from: https://doi.org/10.1038/sdata.2016.18

[2] Health Data Research UK. Data Standards Principles. 2020.

[3] Health Level Seven (HL7). FHIR Overview [Internet]. [cited 2021 Jul 26]. Available from: shorturl.at/csCU9

[4] openEHR Foundation. What is openEHR? [Internet]. 2019 [cited 2021 Jun 6]. Available from: http://bit.ly/openehr

[5] IHE International. Cross-Enterprise Document Sharing (XDS.b) [Internet]. IHE ITI Technical Framework (Rev17 Vol1). [cited 2021 Jul 27]. Available from: shorturl.at/enqHQ

[6] National Library of Medicine. Overview of SNOMED CT [Internet]. 2016 [cited 2021 Jul 27]. Available from: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html

[7] Wettstein R, Merzweiler A, Klass M, et al. Using openEHR in XDS.b Environments – Opportunities and Challenges. Stud Health Technol Inform [Internet]. 2020 [cited 2021 Aug 4];272:300–3. Available from: https://doi.org/10.3233/shti200554

[8] National Institutes of Health. Use of the HL7 Fast Healthcare Interoperability Resources (FHIR) for Capturing and Sharing Clinical Data for Research Purposes [Internet]. 2019 [cited 2021 Jul 26]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-150.html

[9] BRIDG, HL7 & FHIR - Biomedical Research Integrated Domain Group [Internet]. National Cancer Institute. 2021 [cited 2021 Jul 26]. Available from: https://bridgmodel.nci.nih.gov/hl7-fhir

[10] Bönisch C. FAIRness of openEHR Archetypes and Templates. 2019;

[11] Jacobsen A, de Miranda Azevedo R, Juty N, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intell. 2020;2(1–2):10–29.

[12] OpenEHR Clinical Knowledge Manager (CKM) [Internet]. Ocean Informatics / Ocean Health Systems; 2019. Available from: https://ckm.openehr.org/ckm/

[13] The Eight Caldicott Principles [Internet]. UK Caldicott Guardian Council. 2020 [cited 2021 Jul 30]. Available from: https://www.ukcgc.uk/manual/principles

[14] Information Governance Alliance. Records Management Code of Practice for Health and Social Care [Internet]. 2016 p. 38, 65. Available from: shorturl.at/fpETY

[15] Metke-Jimenez A, Steel J, Hansen D, et al. Ontoserver: A syndicated terminology server. J Biomed Semantics. 2018;9(1):1–11.

[16] Frexia F, Mascia C, Lianas L, et al. openEHR Is FAIR-Enabling by Design. Stud Health Technol Inform [Internet]. 2021 May 27 [cited 2021 Jul 21];281:113–7. Available from: shorturl.at/nsxT6

[17] Vesteghem C, Brøndum RF, Sønderkær M, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief Bioinform [Internet]. 2020 May 18 [cited 2021 Jul 21];21(3):936–45. Available from: https://doi.org/10.1093/bib/bbz044

[18] HL7 SOA Work Group. HL7 FHIR and FAIR principles (FAIR4FHIR) [Internet]. FHIR for FAIR Implementation Guide. 2021 [cited 2021 Sep 16]. Available from: shorturl.at/pvMQ7

[19] Kush RD, Warzel D, Kush MA, et al. FAIR data sharing: The roles of common data elements and harmonization [Internet]. Vol. 107, Journal of Biomedical Informatics. J Biomed Inform; 2020 [cited 2021 Jul 25]. Available from: https://pubmed.ncbi.nlm.nih.gov/32407878/

[20] FHIR UK Core - NHS Digital [Internet]. NHS Digtial. 2020 [cited 2020 Apr 8]. Available from: https://digital.nhs.uk/services/fhir-uk-core

[21] Willems M. FAIR Data Maturity Model: specification and guidelines [Internet]. 2020 [cited 2021 Sep 15]. Available from: shorturl.at/inqKO

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210833

Challenges and Experiences Extending the cBioPortal for Cancer Genomics to a Molecular Tumor Board Platform

Niklas REIMER[a,1,*], Philipp UNBERATH[b,*], Hauke BUSCH[a], Melanie BÖRRIES[c,d], Patrick METZGER[c], Arsenij USTJANZEW[e], Christopher RENNER[b], Hans-Ulrich PROKOSCH[b] and Jan CHRISTOPH[b,f]

[a] *Group for Medical Systems Biology, Lübeck Institute of Experimental Dermatology, Universität zu Lübeck, Germany*

[b] *Department of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*

[c] *Institute of Medical Bioinformatics and Systems Medicine, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany*

[d] *German Cancer Consortium (DKTK) Partner Site Freiburg and Cancer Research Center (DKFZ), Heidelberg, Germany*

[e] *Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), Johannes Gutenberg-University School of Medicine, Mainz, Germany*

[f] *Junior Research Group (Bio-)Medical Data Science, Martin-Luther-University Halle-Wittenberg, Faculty of Medicine, Halle, Germany*

Abstract. In Molecular Tumor Boards (MTBs), therapy recommendations for cancer patients are discussed. To aid decision-making based on the patient's molecular profile, the research platform cBioPortal was extended based on users' requirements. Additionally, a comprehensive dockerized workflow was developed to support the deployment of cBioPortal and connected services. In this work, we present the challenges and experiences of nearly two years of implementing and deploying an MTB platform based on cBioPortal and compare those to findings of a previous study.

Keywords. cBioPortal, Molecular Tumor Board, MIRACUM, Precision Medicine, Genomics, Docker

1. Introduction

Next-generation sequencing (NGS) techniques are becoming widespread in personalized cancer treatment [1]. These data are fundamentally important in the context of Molecular Tumor Boards (MTBs), where experts from different fields, like oncology, bioinformatics, and systems medicine, jointly discuss therapy options for cancer patients

[1] Corresponding Author, Niklas Reimer, Group for Medical Systems Biology, Lübeck Institute of Experimental Dermatology, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany; E-mail: n.reimer@uni-luebeck.de.

[*] These authors contributed equally to this work.

based on molecular data. While these advancements have demonstrated their potential to improve patient outcomes already [2,3], the management, analysis, and interpretation of these data poses a challenge to traditional healthcare systems. In research, however, several software tools supporting data processing and interpretation exist, one of which is the cBioPortal for Cancer Genomics [4,5] developed by the Memorial Sloan Kettering Cancer Center (MSKCC). In order to evaluate the impact of bringing such tools to MTBs and personalized cancer treatment, the MIRACUM Use Case 3 [6] is developing a comprehensive workflow and tool architecture from the sequencer to the clinician providing care to cancer patients. One crucial part of this goal is to deploy an MTB software platform based on cBioPortal and extended with various additional functionalities identified from extensive requirements analysis with real-world users in the clinics [7].

2. Objectives

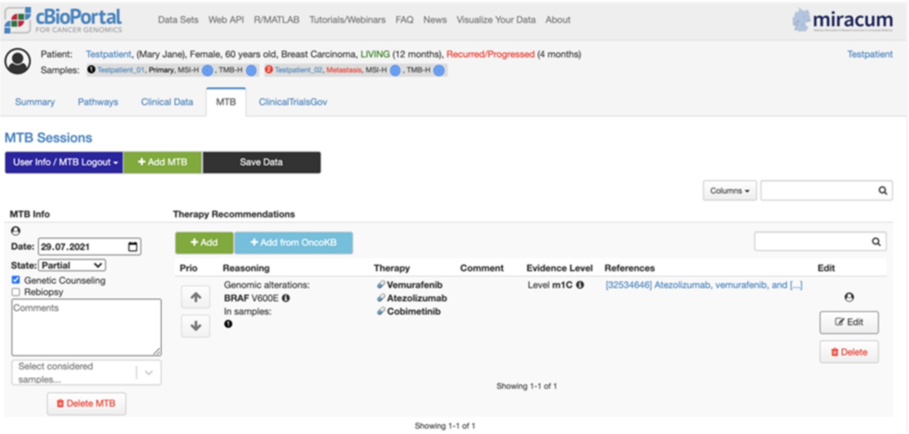
In this work, we share our experience in extending cBioPortal for use in a clinical setting. This includes the challenges and possible solutions encountered and the cBioPortal extensions deployed together with their architectural details, surrounding tools, and milestone releases' rollout. These developments are compared and evaluated concerning the findings of a previous study from 2018 that described extending cBioPortal in a research setting [8] regarding the heterogeneity of system environments and different levels of integrated data available. The deployment processes are monitored through feedback forms to collect issues encountered during the setup.

3. Development

Having gained the first experience with extending cBioPortal (version 1.11.3) in a research setting in 2018 [8], the development of additional functionalities started in late 2019 with the cBioPortal version 3.1.2, derived from a detailed requirements analysis for the use of cBioPortal as an MTB platform [7]. Since it was expected that not all extensions were suitable for contribution to the main cBioPortal project, the development was carried out on a separate repository, forked from the original codebase but kept up to date alongside updates to the cBioPortal codebase [9]. Forking was only necessary for the cBioPortal frontend project as none of the implemented functionalities needed adaptions to the cBioPortal backend or database. During the 22 months of development, a total of 45 updates releases were also transferred to our forked version. All issues during such updates were caused by refactoring in the cBioPortal codebase or moving reusable parts of the codebase into separate packages. Apart from that, about 15 times, merge conflicts occurred but were easy to solve. This means that updates can usually be applied with little or no effort, and long-term support for features implemented in the fork would indeed be possible.

The most significant extensions were two new tabs in the patient view of cBioPortal. Firstly, one to allow for the search of clinical trials based on genomic and clinical data of the patient achieved with direct integration of ClinicalTrials.gov (publication currently under revision). Secondly, one to enable structured and standardized documentation of therapy recommendations of the MTB as shown in Figure 1. This also included a novel and more detailed authentication and authorization concept and the integration of the service FhirSpark to provide a FHIR-compliant way to store therapy recommendations [10]. Both extensions were developed using a user-centered design process and evaluated through a usability test.

There were also several minor extensions and adaptions, e.g., the integration of approval status of drugs by the European Medicines Agency directly in the OncoKB annotation already available in cBioPortal, import and display of LoH-mutations, and handling of internal PDF documents by rendering them through the web browser instead of depending on external Google services.



Patient: Testpatient, (Mary Jane), Female, 60 years old, Breast Carcinoma, LIVING (12 months), Recurred/Progressed (4 months) Testpatient
Samples: 1 Testpatient_01, Primary, MSI-H, TMB-H 2 Testpatient_02, Metastasis, MSI-H, TMB-H

MTB Sessions

MTB Info

Date: 29.07.2021
State: Partial
☑ Genetic Counseling
☐ Rebiopsy
Comments
Select considered samples...

Therapy Recommendations

Prio	Reasoning	Therapy	Comment	Evidence Level	References	Edit
↑ ↓	Genomic alterations: BRAF V600E In samples: 1	Vemurafenib Atezolizumab Cobimetinib		Level m1C	[32534646] Atezolizumab, vemurafenib, and [...]	Edit Delete

Showing 1-1 of 1

Showing 1-1 of 1

Figure 1. The newly implemented MTB tab offers structured documentation of therapy recommendations by selecting clinical and mutation data for reasoning, drugs as a therapy, and the corresponding evidence level based on approval state and available references. Multiple therapy recommendations can be prioritized.

4. Deployment

A major challenge in software deployment was enabling cBioPortal and its connected tools and services, like databases and annotation tools, to be distributed as simply as possible. The official cBioPortal GitHub project site initially provided a solution for the deployment of cBioPortal via Docker [11], requiring manual intervention at multiple points, including a manual setup of the containers. Therefore, we developed a custom workflow [12] for the standardized and simplified deployment of our extended version of cBioPortal, including a MySQL database, session service, an on-premise instance of Genome Nexus [13], and the FhirSpark service described in section 3. This solution also includes the option to deploy the standard cBioPortal in a research setting, which was later used by the team of MSKCC as the basis for the official cBioPortal Docker project [14].

Our Docker-based workflow successfully enabled the deployment of cBioPortal at all ten consortial sites of MIRACUM, along with other partner sites from the HiGHmed [15] and SMITH [16] consortia, as well as sites from the Bavarian Centre for Cancer Research (BZKF) and the German Cancer Consortium (DKTK). The deployment at the consortial sites was accompanied by structured feedback forms to collect comments and problems during the installation. The evaluation of the forms revealed that some sites delegated the deployment to the IT department, while at other sites, domain experts handle the deployment. The only notable difficulties were caused by the integration with other tools, like the configuration of the identity provider Keycloak [17] or site-specific barriers like highly restrictive firewalls, incomplete proxy configurations, or issues integrating Transport Layer Security encryption.

As MIRACUM Use Case 3 intends to provide a comprehensive workflow for MTB case preparation, the deployment project builds upon MIRACUM-Pipe [18]. This sequencing pipeline includes advanced variant annotation and generates an interactive PDF report and files needed for importing mutation data in cBioPortal. Using this complete setup, the whole process from the sequencer to visualization in cBioPortal could be successfully tested at three sites with over 370 whole-exome and panel sequencing cases.

Although extensive documentation of the import data format of cBioPortal is readily available, the generation of the required files to import both clinical and molecular data is challenging. Especially when adding new patients to an existing study, file management becomes prone to error and hardly practicable by hand because IDs must be consistent across patient, sample, and case-list files. At the same time, newly added columns must also be populated in existing patients. To aid this process, cpbManager [19] was developed and integrated into the dockerized deployment workflow. The tool features a user interface that allows for easy import file generation and management.

5. Discussion and Conclusion

The major challenge on the development side is the long-term support of features like the therapy recommendation that are specific to the described use case as they require a regular merging with the upstream codebase. Even though the codebase structure is significantly improving, like refactoring code into separate reusable packages and separating the backend and frontend project [8], this step requires programming skills and thorough testing. However, a plugin concept in cBioPortal would help overcome this issue, depending on its customization capabilities.

Due to the varying delegation of the deployment mentioned in section 4 and complex domain-specific features, it will be necessary to provide detailed documentation. Especially edge cases like different types of proxy configurations were added to the Docker Compose solution, including test data sets for easier verification of proper functionality.

Even though the development of cbpManager lowers the barrier when managing studies with cBioPortal, it still requires manual curation of data that will also be available in the local data integration centers. Therefore, long-term solutions should use these as primary data sources for automated ETL (extract, transform, load) workflows, requiring less manual interaction.

Currently, all provided tools are intended for research use only and are not approved for the diagnosis or treatment of individual patients. However, future efforts in the project will also cover the tools' compliance within the scope of the Medical and In Vitro Diagnostic Regulation (MDR and IVDR).

This work demonstrates how cBioPortal can be extended and integrated with other tools to a comprehensive and easily deployable MTB software solution. While the concurrent development and continuous updates of a forked cBioPortal are not trivial, refactoring the original project significantly impacted the maintainability. However, contributing as many features as possible to the main project should remain the primary goal.

Acknowledgment

The project is funded by the German Federal Ministry of Education and Research (BMBF), Grand IDs 01ZZ1802Z (HiGHmed), 01ZZ1803N (SMITH), 01ZZ1801A, 01ZZ1801B, and 01ZZ1801F (all MIRACUM). We especially thank Nikolaus Schultz and Jianjiong Gao from the MSKCC, who provided us with feedback on the identified requirements and our planned changes to cBioPortal. The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (PU).

References

[1] Garraway LA, Verweij J, Ballman KV. Precision Oncology: An Overview. JCO 2013;31:1803–5.

[2] Hoefflin R, Geißler A-L, Fritsch R, Claus R, Wehrle J, Metzger P, et al. Personalized Clinical Decision Making Through Implementation of a Molecular Tumor Board: A German Single-Center Experience. JCO Precision Oncology 2018:1–16.

[3] Hoefflin R, Lazarou A, Hess ME, Reiser M, Wehrle J, Metzger P, et al. Transitioning the Molecular Tumor Board from Proof of Concept to Clinical Routine: A German Single-Center Analysis. Cancers 2021;13:1151.

[4] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. Cancer Discovery 2012;2:401–4.

[5] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science Signaling 2013;6:pl1–pl1.

[6] Prokosch H-U, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine: A Large Data Sharing Network to Enhance Translational Research and Medical Care. Methods Inf Med 2018;57:e82–91.

[7] Buechner P, Hinderer M, Unberath P, Metzger P, Boeker M, Acker T, et al. Requirements Analysis and Specification for a Molecular Tumor Board Platform Based on cBioPortal. Diagnostics 2020;10:93.

[8] Unberath P, Knell C, Prokosch H-U, Christoph J. Developing New Analysis Functions for a Translational Research Platform: Extending the cBioPortal for Cancer Genomics. Studies in Health Technology and Informatics 2019;258:46–50.

[9] nr23730. cbioportal-frontend - React Frontend of cBioPortal 2020. https://github.com/nr23730/cbioportal-frontend (accessed July 29, 2021).

[10] Reimer N, Unberath P, Busch H, Ingenerf J. FhirSpark – Implementing a Mediation Layer to Bring FHIR to the cBioPortal for Cancer Genomics. In: Mantas J, Stoicu-Tivadar L, Chronaki C, Hasman A, Weber P, Gallos P, et al., editors. Studies in Health Technology and Informatics, IOS Press; 2021.

[11] Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014;2014:Article 2.

[12] buschlab. MIRACUM-cbioportal 2020. https://github.com/buschlab/MIRACUM-cbioportal (accessed July 29, 2021).

[13] Memorial Sloan Kettering Cancer Center. Genome Nexus - Annotation and Interpretation of Genetic Variants in Cancer n.d. https://www.genomenexus.org/ (accessed July 28, 2021).

[14] Memorial Sloan Kettering Cancer Center. Run cBioPortal using Docker Compose 2020. https://github.com/cBioPortal/cbioportal-docker-compose (accessed July 28, 2021).

[15] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods Inf Med 2018;57:e66–81.

[16] Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart medical information technology for healthcare (SMITH). Methods of Information in Medicine 2018;57:e92–105.

[17] WildFly. Keycloak - Open Source Identity and Access Management n.d. https://www.keycloak.org/ (accessed July 28, 2021).

[18] AG-Boerries. MIRACUM-Pipe 2019. https://github.com/AG-Boerries/MIRACUM-Pipe (accessed July 29, 2021).

[19] Ustjanzew A, Marini F. cbpManager: Generate, manage, and edit data and metadata files suitable for the import in cBioPortal for Cancer Genomics. https://arsenij-ust.github.io/cbpManager/index.html; 2021.

doi:10.3233/SHTI210834

Deep Learning, a Not so Magical Problem Solver: A Case Study with Predicting the Complexity of Breast Cancer Cases

My-Anh LE THIEN[a,1], Akram REDJDAL[a], Jacques BOUAUD[b,a] and Brigitte SEROUSSI[a,c,d]

[a] *Sorbonne Université, Université Sorbonne Paris Nord, Inserm, UMR S_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c] *AP-HP, Hôpital Tenon, Paris, France*
[d] *APREC, Paris, France*

Abstract. Using guideline-based clinical decision support systems (CDSSs) has improved clinical practice, especially during multidisciplinary tumour boards (MTBs) in cancer patient management. However, MTBs have been reported to be overcrowded, with limited time to discuss all cases. Complex breast cancer cases that need further MTB discussions should have priority in the organization of MTBs. In order to optimize MTB workflow, we attempted to predict complex cases defined as non-compliant cases despite the use of the decision support system OncoDoc. After previously obtaining insufficient performance with machine learning algorithms, we tested Multi Layer Perceptron for classification, compared various samplers to compensate data imbalance combined with cross-validation, and optimized all models with hyperparameter tuning and feature selection with no improvement and lacklustre results (F1-score: 31.4%).

Keywords. Clinical decision support systems, Deep learning, Breast cancer

1. Introduction

Patient-specific treatment plans as recommended by clinical practice guidelines (CPGs) have improved patient outcomes and clinicians are highly encouraged to implement them [1]. In many countries, the management of cancer patients is discussed in multidisciplinary tumour boards (MTBs) to allow for the best collective decision-making. In addition, clinical decision support systems (CDSSs) have shown to improve the compliance of MTB decisions with CPG recommendations. However, MTBs have been reported to be overloaded with limited time to discuss properly each clinical case.

OncoDoc is a CDSS developed to provide patient-specific recommendations and promote the implementation of breast cancer CPGs [2]. The system has been routinely used in MTBs at the Tenon hospital (Paris, France) in a three-year period. MTB decision compliance rate with CPGs for invasive breast cancers reached 91.7%. In the remaining 8.3%, patients presented specific medical circumstances not formally

[1] Corresponding author, My-Anh Le Thien, Email: myanh.lethien@gmail.com

covered by the CPGs in OncoDoc knowledge base, which explain that clinicians might in some cases disagree with OncoDoc treatment plans and choose not to follow them.

We wish to pre-emptively identify patients whose profiles are too complex to be properly handled by CPGs and optimize patient triage ahead of MTBs. In this way, non-complex cases might be treated more rapidly, and additional time and focus could be allocated to cases identified as complex. As OncoDoc is directly built from CPGs, we hypothesize that non-compliance with OncoDoc recommendations is a marker of case complexity and that predicting cases leading to non-compliant MTB decisions might help identify complex cases.

In a previous study, we used different machine learning algorithms (RandomForest, DecisionTree, XGBoost) to classify OncoDoc's complex cases based on available routine data and tested numerous sampling methods to compensate data imbalance, with unsatisfying results (F1-score did not exceed 40%) [3]. As deep learning is known to offer possible improvement on imbalanced samples when machine learning lacks sufficient performance, we implemented a Multi Layer Perceptron (MLP) on OncoDoc data in order to improve complex cases classification and identification.

2. Material and Methods

2.1. Dataset

Data was collected from the existing OncoDoc database and included MTB decisions for adult women treated for breast cancer from February 2007 to September 2009 at the Tenon hospital (AP-HP, Paris, France). Data consisted of 1,887 MTB decision instances (1,054 patients) with 127 collected variables. A sizable number of variables was incomplete due to OncoDoc's architecture as a decision tree: data not relevant to the case is not asked, and therefore not entered.

We applied supervised deep learning with labelled training datasets and all values predicted from the test datasets verified against the actual class. According to our hypothesis, cases where clinicians did not comply with OncoDoc recommendations (8.3%) were labelled as "complex" whereas others cases were labelled as "non-complex". Whenever possible, missing values were imputed so as to remain logically sound, e.g., when a tumour was non-invasive, all tumour-invasive-related variables were filled as "not applicable". Continuous variables were converted to categorical variables (e.g. patient age) and all variables with labels were encoded as integers. All missing or "not applicable" values were considered as integers, e.g., excision margins outlying invasive tumour originally coded as invaded (1) or not (0) were encoded as missing (0), not applicable to non-invasive tumour (1), non-invaded (2) and invaded (3). Additional variables were built to reflect known factors of clinical complexity, e.g., triple negative breast cancer patients (hormonal receptors = negative AND Her2 receptors = negative). The final dataset consisted of 1,887 decisions and 70 variables.

2.2. Multi Layer Perceptron

MLP is a subtype of artificial neural network (ANN) which uses back-propagation, and presents at least three neuron layers for data classification: an input layer, a hidden layer, and an output layer. In MLP, data is handled one way and passes all layers once, as opposed to the way data is handled in recurrent neural networks. Back-propagation allows estimating what constitutes erroneous values in the hidden layer (which has no reference in the initial data to compare itself to) from the final classification labels [4].

2.3. Splitting and sampling dataset

The following standard procedures were applied for this study:

- With stratification on case complexity: stratification keeps the class ratio between the original dataset and the training and testing datasets.
- With or without k-fold cross-validation: in k-fold cross-validation, the training set is split into k smaller sets (here, k=5). Each of the k “folds” are used in turn as testing set against the rest and cross-validation then averages performance measures to give a more accurate estimate of model performance.

Since data was severely imbalanced, as there were few complex cases, we systematically tested and compared the following samplers on each training datasets generated for cross-validation, in order to offset data imbalance [5]:

- Random Under Sampling (RUS)
- Random Over Sampling (ROS)
- Adaptive Synthetic (ADASYN)
- Synthetic Minority Oversampling Technique (SMOTE)
- SMOTE and Edited Nearest Neighbours (SMOTEEN)

2.4. Hyperparameter tuning and feature selection

We searched the best values for the model’s hyperparameters using Random Search and Grid Search. Random Search tested random combinations of hyperparameters in a given range and Grid Search was applied to narrow down the best values of hyperparameters. We optimized model variables with feature selection: all variables were first included in the analyses, then progressively excluded from the model from lowest to highest feature importance, e.g. from least to most useful variable as identified by the trained model. The optimized number of variables to include was then retained.

2.5. Evaluation indices

MLP was evaluated using precision, recall, and F1-score. Accuracy was available but considered to be unreliable as the testing dataset was unbalanced (if 90% of data belongs to class A, a model might simply choose to systematically class data as class A to obtain a 90% accuracy). We compared the mean cross-validation indices for each sampling methods and model. The overall process is illustrated in Figure 1.

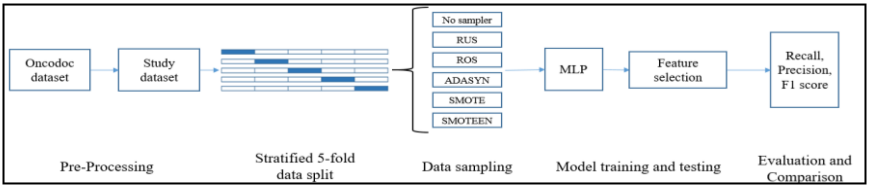


Figure 1. Analytic plan.

3. Results

Before feature selection, MLP presented its best F1 score without any sampler (recall=25.9%, precision=40.4%, F1-score=31.4%). Applying samplers improved recall at the cost of precision, especially for RUS (recall=65.5%, precision=16.9%, F1-score=26.7%) and SMOTEEN (recall=52.3%, precision=20.2%, F1-score=28.6%), with decreased F1-score.

Feature selection did not improve MLP when no sampler was applied but showed slight improvement for ROS (before feature selection: recall=39.1%, precision=19.3%, F1-score=25.6% *versus* after feature selection: recall=47.6%, precision=23.5%, F1-score=30.9%) and ADASYN (before feature selection: recall=33.3%, precision=21.3%, F1-score=25.8% *versus* after feature selection: recall=37.3%, precision=25.0%, F1-score=29.2%). Almost no improvement was observed for SMOTE, and performance stayed the same for SMOTEEN. (cf. Table 1).

Table 1. MLP model scores by sampling technique before and after feature selection (%)

Model	Score	No sampling	RUS	ROS	ADASYN	SMOTE	SMOTEEN
Before feature selection	Recall	25.9	65.5	39.1	33.3	33.3	52.3
	Precision	40.4	16.9	19.3	21.3	26.2	20.2
	F1	31.4	26.7	25.6	25.8	28.6	29.0
After feature selection	Recall	25.9	63.2	47.6	37.3	33.9	50.0
	Precision	40.4	17.0	23.5	25.0	27.5	20.4
	F1	31.4	26.9	30.9	29.2	29.5	28.7

4. Discussion

Multiple sampling methods were used to compensate data imbalance with MLP. Improvement was observed for recall in all samplers but showed deteriorated values for precision, with unsatisfactory F1-score results. In other words, our models could not easily identify complex cases and could only improve recall performance by indiscriminately selecting cases in the hope to identify complex cases, decreasing precision performance. Calculated from recall and precision, F1-score could not increase when one score improved at the expense of the other, and showed the overall performance plateau. Hyperparameters tuning improved all models through Random Search and Grid Search but remains insufficient. Throughout hyperparameters tuning, all models had the tendency to sacrifice precision for recall. Feature selection granted slight improvement by excluding variables presenting no added value to the model. Few variables were consistently more “important” than others e.g., ‘Profile frequency’ or ‘Age_35_75’ are easily explained as known factors of clinical complexity.

As a pre-constructed model in Python, MLP generated results similar or worse than those obtained with machine learning algorithms [3]. However, hyperparameter tuning with MLP was a non-compressible time-consuming task that did not allow for extensive testing and a larger range of parameters could possibly yield better results. On the other hand, we attempted to construct deep learning models “from scratch” without using a pre-existing Python library such as MLP, but all of them choose to classify all data as “non-complex” with no attempt to identify “complex” cases. Such behaviour could not be explained as a way to exploit the data imbalance (when 90% of data was ‘non-complex’) since similar results were obtained after compensating with data sampling. As it is possible that further testing could lead to a functional model “from scratch”, further research and understanding of deep learning models and neural networks might improve our current results.

Errors in data were encountered during pre-processing, faulty recording of patients’ characteristics might have occurred. Likewise, data structure and variables were significantly modified during pre-processing (various variables were outright excluded from analysis, or heavily modified to simplify or avoid missing data) which may have removed important information for data classification. Additional data might also benefit model training and it is plausible that OncoDoc available data simply does not include the information needed to accurately predict complex cases.

Lastly, we assumed non-conformity with OncoDoc as the sole explicit marker of complexity and acknowledge the possible limits associated: complex cases which eventually remained conform to OncoDoc after a lengthy discussion are not identified in our study, likewise, some treatments might be dismissed without posing difficulty to MTB clinicians (such as patient’s preference). Further reviews of MTB decisions and analyses of causes for non-compliance might give us a more accurate indication of complexity for further studies.

5. Conclusion

We collected data from OncoDoc routine data to identify complex patient cases, with deep learning methods, assuming complex cases could not be systematically managed by a CPG-based CDSS. The dataset was prepared for analysis and underwent various sampling techniques to overcome its class imbalance. Several improvement plans were implemented but results did not improve our previous performance with machine learning models, hinting at a possible need for different models, additional data and/or additional data structuring for further improvement.

References

[1] Kreienberg R, Wöckel A, Wischnewsky M. Highly significant improvement in guideline adherence, relapse-free and overall survival in breast cancer patients when treated at certified breast cancer centres: An evaluation of 8323 patients. The breast. 2018 Aug 1;40:54-9.

[2] Séroussi B, Bouaud J, Gligorov J, Uzan S. Supporting multidisciplinary staff meetings for guideline-based breast cancer management: a study with OncoDoc2. InAMIA Annual Symposium Proceedings 2007 (Vol. 2007, p. 656). American Medical Informatics Association.

[3] Le Thien M-A, Redjdal A, Bouaud J, et al. Using Machine Learning on Imbalanced Guideline Compliance Data to Optimize Multidisciplinary Tumour Board Decision Making for the Management of Breast Cancer Patients Studies in Health Technology and Informatics. Stud Health Technol Inform. 2021, October 2-4, To appear

[4] Moghaddasi H, Ahmadzadeh B, Rabiei R, et al. (2017 [cited 2021 Sep 14]) Study on the Efficiency of a Multi-layer Perceptron Neural Network Based on the Number of Hidden Layers and Nodes for Diagnosing Coronary- Artery Disease Jentashapir J Cell Mol Biol. 2017; 8(3):e63032.

[5] Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review Int J Patt Recogn Artif Intell. 23:687–719.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210835

Influence of Healthcare Organization Factors on Cardiovascular Diseases Mortality

Oleg METSKER[a,1] and Georgy KOPANITSA[b]
[a]*Almazov National Medical Research Centre, Saint Petersburg, Russia*
[b] *ITMO University, Saint Petersburg, Russia*

Abstract. One serious pandemic can nullify years of efforts to extend life expectancy and reduce disability. The coronavirus pandemic has been a perturbing factor that has provided an opportunity to assess not only the effectiveness of health systems for cardio-vascular diseases (CVD), but also their sustainability. The goal of our research is to analyze the influence of public health factors on the mortality from circulatory diseases using machine learning methods. We analysed a very large dataset that consisted of the information collected from the national registers in Russia. We included data from 2015 to 2021. It included 340 factors that characterize organization of healthcare in Russia. The resulting area under receiver operating characteristic curve (AUC of ROC) of the Random Forest based regression model was 92% with a testing dataset. The models allow for automated retraining as time passes and epidemiological and other situations change. They also allow additional characteristics of regions and health care organizations to be added to existing training datasets depending on the target. The developed models allow the calculation of the probability of the target for 6-12 months with an error of 8%. Moreover, the models allow to calculate scenarios and the value of the target indicator when other indicators of the region change.

Keywords. Cardiovascular disease, machine learning, public health, prediction, keyword

1. Introduction

One serious pandemic can nullify years of efforts to extend life expectancy and reduce disability [1]. The coronavirus pandemic has been a perturbing factor that has provided an opportunity to assess not only the effectiveness of health systems for cardio-vascular diseases (CVD), but also their sustainability [2,3]. The dynamics of total and CVD mortality can be used as a measure of health system resilience. Efficiency and sustainability are different and, in many ways, mutually exclusive, but sustainability is as necessary for sustained positive dynamics as efficiency is for achieving goals [2]. Analyzing the situation with COVID-19 we understand the need to assess the sustainability of health systems in relation to CVD care, a sustainable system will have different characteristics compared to an effective one [4].

[1] Corresponding Author, Georgy Kopanitsa, ITMO University, Saint-Petersburg, Russia; E-mail: georgy.koapnitsa@gmail.com.

In the post-COVID era, we should strive to build balanced, rather than efficient, systems with sufficient resilience[1]. So, when solving the inverse problem of forecasting an indicator/indicator of quality of treatment of a region it is possible to calculate what the characteristics of the region should be today and tomorrow (values quantitatively qualitatively) to get the required indicator the day after tomorrow.

Models and algorithms for the analysis of risk factors and prognosis of mortality in the acute phase of the disease have been developed. Many works apply methods of artificial intelligence. For example [5], considers the identification of ischemic stroke risk factors in conditions of data shortage. Many studies, such as [6–8] use large population databases, including up to 800,000 patients, to predict stroke incidence over 5 years. Despite rather high accuracy: up to 87% correct prediction of ischemic stroke and up to 82% prediction of hemorrhagic stroke, the developed methods based on neural networks and machine of reference vectors do not allow to work in conditions of uncertainty and data gaps in electronic medical histories.

Much attention is paid to treatment planning and prognosis of recovery in the acute phase of the disease. For ischemic stroke, models based on neural networks and support vector method show the best performance [9,10]. The correctness of the models reaches 74% in the best cases, which cannot be considered a satisfactory result. However, the influence of organizational factors on population mortality from CVD has not been considered in detail in the scientific literature.

1.1. Objectives

The goal of our research is to analyze the influence of public health factors on the mortality from circulatory diseases using machine learning methods.

2. Methods

2.1. Dataset

The dataset consisted of the following information collected from the national registers in Russia. We included data from 2015 to 2021. Information about the activities of the organization providing medical care; Information on the number of diseases registered in patients residing in the service area of the medical organization; information on the movement of patients; Information on confirmed cases of death in the following nosologies and their International Statistical Classification of Diseases and Related Health Problems (ICD 10) codess: Diseases of the circulatory system (I00-I99), acute coronary syndrome (I20- I22), Cerebrovascular diseases - Subarachnoid hemorrhage, Intracerebral hemorrhage, Brain infarction, Stroke not specified as hemorrhage or infarction, Congestion and stenosis of the precerebral arteries, Embolisms, Consequences of cerebrovascular disease (I60-I69), Novoplasms (C00-D48) including oncohematological patients with C90, Delivery O80-O84, Endocrine diseases, eating disorders and metabolic disorders (E00-E90) including diabetes and obesity, as well as death from Sepsis (A40-41), Anemias (D50-D64), Selected disorders involving the immune mechanism (D80-D89), Obesity (E66), Chronic rheumatic heart disease (I05-I09), Influenza (J09-J11), Acute respiratory upper respiratory tract infections (J00-J06, line 11. 1)); Information on the staff of medical organizations, information on surgical work, information on resources of clinics. Indicators of socio-economic development of regions.

2.2. Machine learning

The regression task for predicting the CVD mortality was solved using the scikit-learn library. In total 340 indicators were used as predictors. Each experiment ran in the setting of stratified 5-fold cross-validation (i.e., random 80% of records were used for training and 20% for testing, target class ratios in the folds were preserved).

For the performance assessment, we ran it 100 times; and 100 x 5-fold cross-validation with total of 500 predictions. As an additional performance assessment score, we used the AUC of ROC. The AUC was calculated based on an average of 5 curves (one curve per fold in the setting of 5-fold cross-validation). Features importance was calculated using a random-forest model.

3. Results

The resulting AUC of ROC of the Random forest based regression model was 92% with a testing dataset. Figure 1 shows the result of calculating the significance of predictors using a machine learning model in solving the regression problem. Training was performed on the data set of 340 indicators of RF regions from 2015 to 2021, including both dynamic indicators (spread of coronavirus infection, mortality from other nosologies including cerebrovascular diseases, coverage of vaccination campaign, population movement, etc.), intensity and coverage of measures to reduce mortality in the region. Examples of such activities were the number of publications in the media

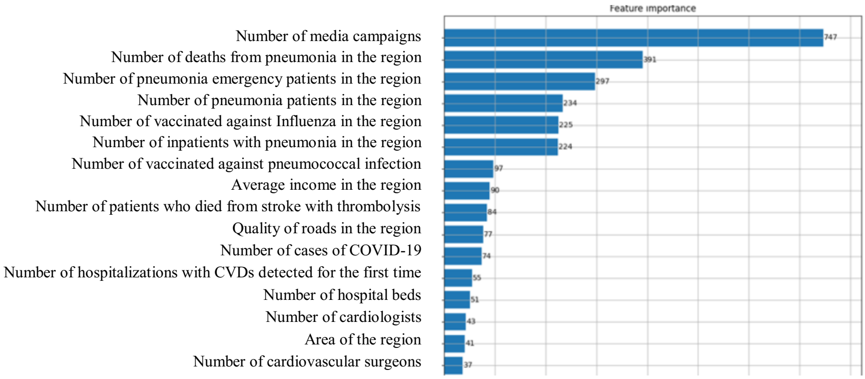


Figure 1. Example of calculating the contribution of regional indicators to CVD mortality using machine learning methods

4. Discussion

The models allow for automated retraining as time passes and epidemiological and other situations change. They also allow additional characteristics of regions and health care organizations to be added to existing training datasets depending on the target. The developed models allow the calculation of the probability of the target for 6-12 months with an error of 8%. Moreover, the models allow to calculate scenarios and the value of the target indicator when other indicators of the region change.

5. Conclusion

The development and implementation of medical information technologies based on machine-learning methods contributes to the development of a unified accessible methodology for analyzing the processes of providing medical care for quality management at all levels of the healthcare system, while maintaining the success achieved in informatization and the existing infrastructure without significant additional costs.

Acknowledgements

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2020-901). This work financially supported by the government of the Russian Federation through the ITMO fellowship and professorship program.

References

[1] Palmer K, Monaco A, et al. The potential long-term impact of the COVID-19 outbreak on patients with non-communicable diseases in Europe: consequences for healthy ageing. Aging clinical and experimental research. 2020 Jul;32:1189-94.

[2] Antony J, Sreedharan R, Chakraborty A, Gunasekaran A. A systematic review of Lean in healthcare: a global prospective. International Journal of Quality & Reliability Management. 2019 Sep 2.

[3] Gasmi A, Peana M, Pivina L, Srinath S, Benahmed AG, Semenova Y, Menzel A, Dadar M, Bjørklund G. Interrelations between COVID-19 and other disorders. Clinical Immunology. 2020 Dec 14:108651.

[4] Huang Y, Cai X, Mai W, Li M, Hu Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. Bmj. 2016 Nov 23;355.

[5] Reberg K, et al. Chronic subdural hematoma, atrial fibrillation and ishemic stroke, considerations about treatment options, a caserepport. Eur Stroke J.2018;3.

[6] Glymour MM, Maselko J, Gilman SE, Patton KK, Avendano M. Depressive symptoms predict incident stroke independently of memory impairments. Neurology. 2010 Dec 7;75(23):2063-70.

[7] Li T, Li G, Guo X, Li Z, Yang J, Sun Y. Predictive value of echocardiographic left atrial size for incident stoke and stroke cause mortality: a population-based study. BMJ open. 2021 Mar 1;11(3):e043595.

[8] Watanabe J, Kakehi E, Kotani K, Kayaba K, Nakamura Y, Ishikawa S. Isolated low levels of high‐density lipoprotein cholesterol and stroke incidence: JMS Cohort Study. Journal of clinical laboratory analysis. 2020 Mar;34(3):e23087.

[9] Wu G, Chen X, Lin J, Wang Y, Yu J. Identification of invisible ischemic stroke in noncontrast CT based on novel two‐stage convolutional neural network model. Medical Physics. 2021 Mar;48(3):1262-75.

[10] Liu Y, Yin B, Cong Y. The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model. Sensors. 2020 Jan;20(17):4995.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210836

Are Semantic Annotators Able to Extract Relevant Complexity-Related Concepts from Clinical Notes?

Akram REDJDAL[a,1], Jacques BOUAUD[a,b], Joseph GLIGOROV[c,d] and Brigitte SÉROUSSI[a,d,e]
[a] *Sorbonne Université, Université Sorbonne Paris* Nord*, Inserm, UMRS_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c]*Sorbonne Université, Institut Universitaire de Cancérologie, Paris, France*
[d]*AP-HP, Hôpital Tenon, Paris, France*
[e]*APREC, Paris, France*

Abstract. Clinical decision support systems (CDSSs) implementing cancer clinical practice guidelines (CPGs) have the potential to improve the compliance of decisions made by multidisciplinary tumor boards (MTB) with CPGs. However, guideline-based CDSSs do not cover complex cases and need time for discussion. We propose to learn how to predict complex cancer cases prior to MTBs from breast cancer patient summaries (BCPSs) resuming clinical notes. BCPSs being unstructured natural language textual documents, we implemented four semantic annotators (ECMT, SIFR, cTAKES, and MetaMap) to assess whether complexity-related concepts could be extracted from clinical notes. On a sample of 24 BCPSs covering 35 complexity reasons, ECMT and MetaMap were the most efficient systems with a performance rate of 60% (21/35) and 49% (17/35), respectively. When using the four annotators in sequence, 69% of complexity reasons were extracted (24/35 reasons).

Keywords. Information Extraction, Decision Support, Breast Cancer

1. Introduction

In many countries, the treatment of cancer patients must be decided in multidisciplinary tumor boards (MTBs). These meetings have been introduced to provide a collaborative and multidisciplinary approach to cancer care, bringing together surgery, oncology, radiology, and pathology specialists to optimize the decision-making process. Prior to MTBs, physicians in charge of patients whose cases will be discussed prepare a breast cancer patient summary (BCPS) as the basis of the oral presentation of the patient case to all MTB clinicians. However, the benefits of MTBs, which have long been taken for granted, are recently being challenged. Positive outcomes from MTBs depend on the presence of qualified and effective faculty, good preparation of patient cases, efficient leadership, sound discussions, and contributive interactions among MTB clinicians [1].

[1] Corresponding author, Akram REDJDAL, Email: redjdalakram300@gmail.com

Clinical decision support systems (CDSSs) are software components that aim to support clinicians in their decision-making process. CDSSs have proven to increase the compliance of clinician decisions with clinical practice guidelines (CPGs) [2]. DESIREE is a European project which aimed at developing web-based services for the management of primary breast cancer by MTBs. During the evaluation of the guideline-based CDSS of DESIREE, we found that for some patient cases the system did not provide any therapeutic proposals or gave recommendations that were not followed by MTB clinicians [3]. These clinical cases were considered as “complex cases”, and we made the assumption that such cases were not correctly handled by guideline-based CDSSs. In the perspective of ultimately building a CDSS able to distinctly support therapeutic decision for complex and non-complex breast cancer cases, the first issue is to identify complex breast cancer cases.

Replicating the mode of operation of MTBs, the aim is to use BCPSs to predict complexity. The first step is to check whether BCPSs do embed complexity-related concepts. As BCPSs are expressed as natural language clinical, non-structured notes, we used different annotators and compared the annotations automatically generated to the reasons of complexity established by a group of clinicians on a sample of BCPSs[2].

2. Material and Methods

2.1. Breast cancer patient summaries

We worked on a sample of 24 BCPSs available as textual unstructured documents. They provide a portrait of patients with all relevant information that MTB clinicians need to know to make the best patient-specific therapeutic decision. BCPSs contain different types of information: reason for presentation, type of tumor, biometric data, personal history, family history, TNM classification, etc. However, unstructured formats make information extraction complicated (e.g., there are many abbreviations, acronyms, and specialized terms.). These 24 BCPSs were manually annotated as “complex” or “non-complex” by a group of seven MTB clinicians of different levels of expertise (from junior to senior) and from different domains (5 oncologists, 2 surgeons). When a clinician considered a clinical case was “complex”, (s)he had to explain why and give the reason of the complexity in terms of patient characteristics.

2.2. Annotation tools: ECMT, SIFR, cTAKES, and MetaMap

We implemented four automatic semantic annotators to extract data from BCPSs. Currently, two systems are widely used in the biomedical field for the English language [4], MetaMap and cTAKES. Since we work on a corpus of French BCPSs, we also considered two systems that work for the French language, i.e., ECMT and SIFR [5].

- MetaMap was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS).

[2] This work has been financed by a doctoral grant for AR from the University Institute of Health Engineering (IUIS, Sorbonne University, Paris, France) and received the support of AP-HP health data warehouse.

The tool uses a hybrid approach combining natural language processing (NLP), knowledge-intensive approach, and computational linguistic techniques.

- cTAKES for Clinical Text Analysis and Knowledge Extraction System uses rule-based and machine learning to extract information from clinical text.
- ECMT (*Extracteur de Concepts Multi-Terminologique* http://ecmt.chu-rouen.fr) is a webservice inspired by the CISMef algorithm for information retrieval with Doc'CISMeF. ECMT works for the French language with seven terminologies and supports semantic expansion features.
- SIFR for Semantic Indexing of French Biomedical Data Resources (http://bioportal.lirmm.fr/annotator) annotator is an openly available web service enabling both recognition and contextualization of concepts from 30 medical terminologies and ontologies.

2.3. Pre-treatment of clinical notes

As cTAKES and MetaMap work on English notes, we translated BCPSs from French to English. However, BCPSs contain a lot of acronyms related to the oncological field (e.g., "*HTA*", "*IRM*", "*TEP*"), difficult to translate with a translator. To solve this issue, we created a local dictionary with medical acronyms and their definition based on online available dictionaries. Then, we replaced acronyms in BCPSs by their definition to get a "translatable" text. We finally used a pre-trained Opus-MT translation model. As a result, all BCPSs were available in French and English in textual format (.txt) used as input by the four annotators. For each system, concepts, CUIs (if available), negation, and certainty were extracted. With ECMT, we used the labels of extracted terms to extract CUIs, but we didn't have information about the context (negativity and certainty) [6].

2.4. Evaluation of annotators

From the corpus of BCPSs, we considered that a BCPS described a complex case if it was considered as complex by *at least* one of the seven MTB clinicians. For each of the complex BCPSs, we collected the list of reasons of complexity as provided by MTB clinicians, and we manually checked whether each element of the list was present in the list of extracted annotations.

3. Results

Among the 24 BCPSs, 14 were considered as complex cases, with seven considered as complex *by all* MTB clinicians. We got 35 reasons of complexity. ECMT and MetaMap were the most efficient systems in terms of complexity parameters extraction, ECMT extracted 60% (21/35) of complexity reasons and MetaMap 49% (17/35). SIFR identified 11 complexity parameters (31%) and cTAKES was the less efficient annotator with only 7 parameters (20%). When using the four annotators in sequence, 24 out of the 35 complexity reasons were extracted (69%). Table 1 shows for each BCPS the reasons of complexity and by which annotator they were retrieved.

Table 1. Evaluation of the four annotators on MTB-clinician-provided complexity-related concepts

BCPS	# MTB clinicians	Reason of complexity	ECMT	SIFR	cTAKES	MetaMap
1	7	Pregnancy	yes	yes	yes	yes
		Patient preference (Refusal of recommended treatment)	no	no	no	yes
		Social situation	yes	no	no	yes
2	7	Radio chemotherapy before surgery	yes	yes	no	yes
		No response to standard treatment	yes	no	no	yes
		Inflammatory syndrome	yes	yes	yes	yes
3	7	Patient preference (Refusal of recommended treatment)	yes	no	no	yes
		Incomplete histology	no	no	no	no
4	7	Comorbidities (age, obesity)	yes	no	no	yes
		Incomplete record	no	no	no	no
		Inadequate margins of excision	yes	yes	no	no
		Use of Oncotype DX	yes	no	no	no
5	7	Comorbidities (age)	yes	yes	yes	yes
		Double cancer	yes	no	no	yes
		Polymedication	yes	yes	yes	yes
6	7	Complex surgical decision	no	no	no	no
7	7	Rare situation	no	no	no	no
		Comorbidities (type 2 diabetes)	yes	yes	yes	yes
		Unclear history of the disease	no	no	no	no
8	6	Prophylactic situation	yes	yes	no	yes
		Family antecedents of breast cancer	yes	no	no	no
		Multifocal cancer	no	no	no	no
9	5	Use of Oncotype DX	yes	no	no	no
		Malignancy	yes	yes	yes	yes
10	5	Incomplete record	no	no	no	no
		Use of Oncotype DX	yes	no	no	no
11	3	Complex surgical decision	no	no	no	no
		Complex surgical decision	yes	no	no	yes
		Multiple imaging procedures needed	no	no	no	no
12	3	Use of Oncotype DX	yes	no	no	no
		Discrepancies between ultrasound and MRI	no	no	no	no
		Multiple metastatic lymph nodes and malignancy	yes	yes	no	yes
13	3	Discrepancies between biopsy and excised tissues	no	no	no	no
		Comorbidities (age)	yes	yes	yes	yes
14	2	Patient preference (Refusal of recommended treatment)	no	no	no	no

4. Discussion and conclusion

We implemented four annotators to assess whether they were able to extract relevant complexity-related concepts from BCPSs. All systems are efficient to extract clear medical concepts (pregnancy, inflammatory syndrome, etc.). ECMT and MetaMap were the most efficient systems as they extracted six parameters that were not extracted by SIFR and cTAKES. ECMT was able to identify two parameters ("Use of Oncotype DX" and "Family antecedents of breast cancer") that were not identified by the other annotators, which can be explained by the fact that ECMT is linked to terminologies that contain these concepts. MetaMap was able to detect one parameter related to patient preference ("Refusal of recommended treatment") that was not extracted by the other annotators. However, this parameter was present in two BCPSs and MetaMap only extracted it once. Three parameters were specifically not extracted by cTAKES, which can be explained by the fact that we used the default clinical pipeline of cTAKES. Indeed, studies reported that other pipelines used for extracting cancer-related information showed good results [7]. It is noticeable that one parameter was only extracted by French annotators ("Inadequate margins of excision"), which may be due to a translation problem. Complexity-related concepts not found by the annotators are context or patient-related parameters, e.g., "Refusal of the recommended treatment", "Complex surgical decision", "Discrepancies between ultrasound and MRI". These parameters are interpreted by clinicians during MTBs but are not explicitly written in BCPSs.

Annotation of BCPSs is time-consuming and labor-intensive for MTB clinicians and automatic semantic annotators when used in sequence may help extracting complexity-related structured concepts from non-structured BCPSs. This would allow us to train machine learning algorithms from automatically generated annotations to categorize complex and non-complex cases ahead of MTBs.

References

[1] El Saghir NS, Keating NL, Carlson RW, Khoury KE, Fallowfield L. Tumor boards: optimizing the structure and improving efficiency of multidisci-plinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ Book. 2014:e461-6. doi: 10.14694/EdBook_AM.2014.34.e461. PMID: 24857140.

[2] Bouaud J, Séroussi B, Antoine EC, et al. A before-after study using OncoDoc, a guideline-based decision support-system on breast cancer management: impact upon physician prescribing behaviour Stud Health Technol Inform. 2001;84:420–424.

[3] Redjdal A, Bouaud J, Guézennec G, Gligorov J, Seroussi B. Reusing Decisions Made with One Decision Support System to Assess a Second Decision Support System: Introducing the Notion of Complex Cases. Stud Health Technol Inform. 2021 May 27;281:649-653. doi: 10.3233/SHTI210251. PMID: 34042656.

[4] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018 Sep 14;18(Suppl 3):74. doi: 10.1186/s12911-018-0654-2. PMID: 30255810; PMCID: PMC6157281.

[5] Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, Metzger MH. Evaluation of a French medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. InMEDINFO 2010 2010 (pp. 252-256). IOS Press.

[6] Redjdal A, Bouaud J, Guézennec G, Gligorov J, Seroussi B. Comparison of MetaMap, cTAKES, SIFR, and ECMT to Annotate Breast Cancer Patient Summaries. Stud Health Technol Inform. 2021, October 2-4, To appear.

[7] Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, Harris D, Hochheiser H, Lin C, Chavan G, Jacobson RS. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res. 2017 Nov 1;77(21):e115-e118. doi: 10.1158/0008-5472.CAN-17-0615. PMID: 29092954; PMCID: PMC5690492.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210837

Chios Hospital Information System Assessment

Konstantinos KARITIS[1], Parisis GALLOS, Ioannis S. TRIANTAFYLLOU and Vassilis PLAGIANAKOS
Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece

Abstract. A very important aspect for organizations that provide healthcare services is to have fully functional and successful information systems. A successful hospital information system can contribute to high quality healthcare services provided to the patients of the hospital. In this paper, is presented the evaluation of the information system of Chios Hospital, "Skylitsio". The survey was conducted using a questionnaire which consists demographic questions and questions that measure the factors of the DeLone & McLean success model. The participants of the survey were 71 users of the clinical information system. Cronbach's alpha reliability test, descriptive statistics, and further data analyses to investigate the relations between the factors of the DeLone & McLean success model were performed. Based on the results, the users of the information system are satisfied with it, as well as they find the system useful and easy to use. The average value of the “information quality” is 3.78 out of 5, the “system quality” is 3.61, the “service quality” is 3.45, the “use” is 3.83, the “user satisfaction” is 3.46, and the “user benefit” is 3.76. The research concludes with a validation of the DeLone & McLean success model and it seems that the information system of the General Hospital of Chios is successful based on the users' opinions.

Keywords. Evaluation, Hospital Information System, Success, DeLone & McLean, “Skylitsio” Chios General Hospital, Information Quality, System Quality, Use, User Satisfaction, Perceived Benefits

1. Introduction

The last few decades, a lot of information and communications technologies were applied in Hospitals and other healthcare organizations to cover administrative and financial needs, as well as to manage patient information [1]. The main goal of these systems was to simplify the communication and documentation through the use of standardized orders as well as, care or treatment plans [2]. A hospital information system (HIS) can manage patients’ admissions, medical records, accounting information, several services, nursing, laboratories, radiology, pharmacy, central procurement, dietary services, staff and payroll data. HIS are integrated computer systems designed to facilitate the management of all medical and administrative data of a hospital and also to improve the quality of the provided healthcare services [3]. The Information systems evaluation is an important process [4], especially on healthcare setting, as it can ensure the efficiency and the

[1] Corresponding Author, Konstantinos Karitis, Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece; E-mail: konstantinoska98@gmail.com.

effectiveness of the HIS. According to the international literature, the success of the information systems plays an important role to the organization's performance [5]. The success of an information system can be measured by three levels, the Organizational level, the Process success level, and the Individual level where the satisfaction from the system usage and the perceived usefulness by the users are examined [5].

Information systems success model (IS success model) was created by DeLone and McLean in 1992 [6] and it is one of the most well-known evaluation models. The model includes the "System Quality" factor which describes how "good" the information system is in terms of its functional characteristics. The "Information Quality" defines how "good" the information system is in terms of its outputs. The "Service quality" describes how "good" the information system is in terms of its available services. The "System Use" refers to the use and utilization of outputs by the information system itself. The "User Satisfaction" measures how satisfied the users are as they use the system and is considered as an important parameter for measuring the success of an information system. The System "Benefits" refers to the benefits that system can offer and is an important aspect of the overall value of the system to its users or organization [7]. The IS success model has been broadly used for information systems assessment in several domains and in healthcare domain too [8-11].

The purpose of this research is to examine the success of the clinical information system of "Skylitsio" General Hospital of Chios using the DeLone and McLean success model.

2. Methods

A paper-based questionnaire was used as a research tool that was distributed among the users of Chios Hospital Information System in June 2020 after a relevant approval of the Hospital's scientific committee. "Skylitsio" General Hospital of Chios is located in the northeastern Aegean Island of Chios, in Greece. The questionnaire aimed to record the opinions of the users of the clinical information subsystem of the hospital. The questionnaire was created based on the DeLone & McLean information systems success model [6-7] and was divided into two parts. The first part included four demographic questions (about age, gender, specialty of the user, computer systems familiarity) while the second part included twenty-two questions to assess six factors of the DeLone & McLean evaluation model [7]. The questions were based on other surveys using the DeLone & McLean evaluation model and were translated in Greek language [8-14]. The factors that where examined were "information quality", "system quality", "service quality", "system usage" ("use"), "user satisfaction", "perceived benefits from using the system" ("benefits"). The aforementioned factors were measured by a five-point Likert scale. The Linkert five-point scale consists of five predefined answers that correspond to a specific numerical value from one to five and express the degree of agreement or disagreement with a particular statement [15]. The data analyses have been performed using SPSS version 25 software. More specifically, the questionnaire's reliability was examined by applying Cronbach's alpha test, as well as new variables were created to investigate the relations between the factors. Descriptive statistics of all the variables of the questionnaire were calculated and correlation analyses between the factors was conducted. To examine the relations between the factors, Spearman correlation coefficients were calculated. Finally, Linear regression analysis was used to evaluate the research model.

3. Results

71 users of the Chios Hospital Information system participated in this survey. From the 71 users, 25 (35.2%) are men and 46 (64.8%) are women. The mean age of the study participants was 42.03, 35 (49.3%) belong to the administrative staff, 21 (29.6%) belong to the Nursing staff, 12 (16.9%) belong to the medical staff and 3 (4.2%) belong to the rest of the hospital staff. The users who consider that they have high familiarity with computers and ICT (Information and Communication Technologies) were 38 (53.5%), those who consider that they have average familiarity with computers and ICT were 31 (43.7%) and 2 (2.8%) consider that they have low familiarity with computers and ICT. The internal consistency of the questionnaire items is reflected in Cronbach's Alpha coefficient which is 0.936. Table 1 presents the average of each dimension, the Cronbach's Alpha scores and the spearman correlation results. All the results found be statistically significant with p-value < 0.01. Multiple linear regression was performed to investigate (a) how "system quality" and "information quality" are associated with "use", (b) how "system quality" and "information quality" are associated with "user satisfaction" and (c) how "use" and "user satisfaction" are associated with "benefits". The linear regression equation for (a) is "use" = 0.538 + (0.328 * "system quality") + (0.577 * "info quality") with p-value < 0.01 and R2 = 0.371. For (b) the linear regression equation is "user satisfaction" = -0.374 + (0.518 * "system quality") + (0.521 * "info quality") with p -value <0.01 and R2 = 0.675. Finally, the linear regression equation for (c) is "system benefits" = 1.46 + (0.401 * "use") + (0.222 * "user satisfaction") with R2= 0.605 and p-value <0.01.

Table 1. Table 1 presents the average of each dimension, the Cronbach's Alpha scores and the spearman correlation results.

Dimensions	Average Value (max=5)	Cronbach's Alpha Coefficient	Associated Dimension	Spearman Correlation Coefficient	p-value
Information Quality	3.78	0.810	Use	0.497	<0.01
			User Satisfaction	0.511	<0.01
System Quality	3.61	0.758	Use	0.424	<0.01
			User Satisfaction	0.487	<0.01
Use	3.83	0.891	User Satisfaction	0.714	<0.01
			Benefits	0.467	<0.01
User Satisfaction	3.46	0.913	Benefits	0.411	<0.01
Benefits	3.76	0.878	User Satisfaction	0.411	<0.01
			Use	0.467	<0.01
Service Quality	3.45	0.767			

4. Discussion

The average value for all the variables and the model factors is above the median (2.5 out of 5), so it can be assumed that the users have a positive opinion about the information system quality, data quality, and services quality. The system usage, the user satisfaction and the perceived benefits of the system usage are also above the median and can be assumed as positive results. Most users consider the information system to have high "information quality", high "system quality" and high "service quality". In particular, they believe that the information receiving by the system is correct, useful, accurate and reliable, they also consider the information system to be easy to use, flexible and easily learnable. Users also consider that they can rely on an information system to get the information they need. They also believe that there is adequate infrastructure and technical support for the system. Users also find the information system useful as it improves the performance of their work, facilitates their work and helps them complete their tasks faster. They also believe that they benefit from using the system as the information system facilitates access to patient information, improves patient care, helps make better decisions and helps to create a paperless environment, only through electronic means. According to the aforementioned results, it seems that the quality of the system affects the users' satisfaction. The better they consider the system to be, the more satisfied they are with it. It also seems that the quality of information affects the use of the system by the Hospital staff. The higher the quality of the information is, the more users use the system. In addition, it seems that the quality of information also affects the satisfaction of the use of the system. The higher the quality they consider the information produced by the system, the more satisfied they are with its use. Furthermore, it seems that the use of the system affects the satisfaction felt by the staff of the Hospital. The more they use the system the more satisfied they feel with it. Supplementary, it seems that the use of the system affects the benefit felt by the staff of the Hospital. The more they use the system the more they feel they receive benefits from it, also it seems that the satisfaction from the use of the system affects the benefit that the staff of the Hospital considers having from its usage. The more satisfied they are with the use of the system, the more they feel they benefit from it.

This research also found that the value of the variable of the factor "use" is affected by changes in the value of the factor "system quality" but even more than the changes of the value of the factor "information quality", also the value of the variable of the factor "user satisfaction" is affected by changes in the value of the factor "information quality" but even more than the changes in the value of the factor "system quality" and the value of the variable factor "benefits from the use of the system" is affected by changes in the value of the factor "use" but not from changes in the value of the "user satisfaction" factor. Based on the above, it can be concluded that there are statistically significant relations between some of the factors of the DeLone & McLean IS success model. Other previous related studies [8-10;16-20] have been produced similar results regarding the success of Hospital Information Systems and other Healthcare Information Systems around the globe.

5. Conclusions

Evaluating the success of an information system is an important and necessary process because through the assessment shortcomings and errors of the system can be identified. Thus, the system evaluation contributes to the system's smooth operation and to the quality of the provided health services. In the present study, the success of the clinical information system of Chios General Hospital "Skylitsio" was investigated using the DeLone & McLean model. According the aforementioned positive, it seems that the information system of Chios Hospital is successful based on the opinions of its users. The research had some limitations, like the large workload of the staff of the Hospital but also the high Covid-19 cases in the island of Chios during the research period that made the data collection very difficult and the sample size quite small. Future work includes a proposal for system upgrade based on evaluation results and needs assessment and the evaluation of other subsystems of the Hospital Information System based on more users' opinions.

Acknowledgements

Authors would like to thank the staff and the scientific committee of "Skylitsio" Chios General Hospital for their cooperation.

References

[1] Hammond WE. Hospital information systems: a review in perspective. Yearbook of medical informatics. 1994; 3.01: 95-102.

[2] Ozbolt G, Bakken S. 'Patient care systems', In: Shortliffe, E. and Perreault, L. (ed.), Medical Informatics: Computer Applications in Health Care and Biomedicine, 2nd ed., New York: Springer. 2001;421-422.

[3] Venot, A, Burgun A and Quantin C. Medical Informatics, e-Health, Verlag France: Springer. 2014.

[4] Hirschheim R, Smithson S. 'Evaluation of information systems: A critical assessment'. In Willcocks P. and Lester S. (ed.), *Beyond the IT Productivity Paradox*, (pp. last chapter), Chichester, UK: John Wiley & Sons. 1999.

[5] Garrity E, Sanders L. Information-systems-success-measurement, Pennsylvania: IGI Global. 1998.

[6] DeLone WH, McLean ER. Information systems success: the quest for the dependent variable. Journal of Information Systems Research. 1992; 3.1:60-95.

[7] DeLone WH and McLean ER. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems. 2003;19.4:9–30.

[8] Ojo AI. Validation of the delone and mclean information systems success model. Healthc Inform Res. 2017;23(1):60–6.

[9] Ebnehoseini Z, Tabesh H, Deldar K, Mostafavi SM, Tara M. Determining the Hospital Information System (HIS) Success Rate: Development of a New Instrument and Case Study. Open Access Maced J Med Sci. 2019 May 14;7(9):1407-1414.

[10] Arana L, Medina L, Pol E, Tun N. Measuring the Success of Hospital Information System (HIS) at La Loma Luz Adventist Hospital. Available from: https://ojs.ub.edu.bz/index.php/PRNDC/article/view/261/105

[11] Thanos L, Gallos P, Zoulias E, Mantas J. Investigating the Success of "Asklepieio Voulas" Hospital Information System. Stud Health Technol Inform. 2021 May 27;281:620-624.

[12] Gallos P, Minou J, Routsis F, Mantas J. Investigating the Perceived Innovation of the Big Data Technology in Healthcare. Stud Health Technol Inform. 2017;238:151-153.

[13] Gallos P, Daskalakis S, Katharaki M, Liaskos J, Mantas J. How do nursing students perceive the notion of EHR? an empirical investigation. Stud Health Technol Inform. 2011;169:243-7.

[14] Stylianides A, Mantas J, Roupa Z, Yamasaki EN. Development of an Evaluation Framework for Health Information Systems (DIPSA). Acta Inform Med. 2018;26(4):230-234.

[15] IS Research Wiki (2019). Semantic theory of survey response. Obtained through the Internet: https://is.theorizeit.org/wiki/Semantic_theory_of_survey_response, [accessed 8/5/2020].

[16] Alipour J, Karimi A, Ebrahimi S, Ansari F, Mehdipour Y. Success or failure of hospital information systems of public hospitals affiliated with Zahedan University of Medical Sciences: A cross sectional study in the Southeast of Iran. Int J Med Inform 2017;108(August):49–54.

[17] Elsadig M, Nassar DA, Menzli LJ. Healthcare Information System Assessment Case Study Riyadh's Hospitals-KSA. In: First International Conference on Computing, ICC 2019 Proceedings, Part II [Internet]. 2019; 252–62.

[18] Saghaeiannejad-Isfahani S, Saeedbakhsh S, Jahanbakhsh M, Habibi M. Analysis of the quality of hospital information systems in Isfahan teaching hospitals based on the DeLone and McLean model. J Educ Health Promot. 2015;4(February):5.

[19] Cho KW, Bae SK, Ryu JH, Kim KN, An CH, Chae YM. Performance evaluation of public hospital information systems by the information system success model. Healthc Inform Res. 2015;21(1):43–8

[20] Mamma E. Evaluation and Quality of Information Systems in Institutional Organizations. 17th Hell ConfAcabemicLibr [Internet]. 2008; p. 1–14. Ibrahim R, Auliaputra B, Yusoff RCM, Maarop N, Zainuddin NMM, Bahari R. Measuring the Success of Healthcare Information System in Malaysia: A Case Study. IOSR J Bus Manag. 2016;18(4):100–6.

doi:10.3233/SHTI210838

Changes in Users Trends Before and During the COVID-19 Pandemic on WHO's Online Learning Platform

Heini UTUNEN[a,1] , Ngouille NDIAYE[a], Lama MATTAR[a], Paula CHRISTEN[a],
Oliver STUCKE [a] and Gaya GAMHEWAGE [a]
[a] *World Health Organization, Geneva, Switzerland*

Abstract. OpenWHO provides open access, online, free and real time learning responses to health emergencies. Before the pandemic, courses on 18 diseases were provided. The increase to 38 courses in response to COVID-19 have led to a massive increase in the number of new learners. As a result, the COVID-19 pandemic affected learners' trends. This paper presents initial findings of changes perceived in the use and user groups' attendance to the World Health Organization's (WHO) health emergency learning platform OpenWHO. Enrolment statistics were based on data collected in December 2019 and March 2021. A descriptive analysis was conducted to explore changes in the usage pattern of the platform. Several user characteristics shifted between before and during the pandemic. More women, younger and older learners joined the learning during the pandemic. Public health education leaned toward a more equitable reach including previously underrepresented groups.

Keywords. Online Learning, COVID-19, OpenWHO, Public Health Emergencies

1. Introduction

OpenWHO provided 18 different disease courses before the COVID-19 pandemic, and 38 courses for COVID-19 response have led to a massive increase in the number of new learners from 140,000 users in 2019, to 5,100,000 by 31 March 2021. A significant change in use by women users, geographical location, and user characteristics. The pandemic, along with the consequent imposed social distancing boosted online learning [1]. This paper presents initial findings of changes perceived in the use and user groups' attendance to OpenWHO.

2. Methods

The enrolment data statistics were drawn from OpenWHO's built-in reporting system, which tracks learners' enrolments, completion rates, demographics and other key course-related data and later processed and disaggregated using Microsoft® Power BI tool to analyze changes in the usage pattern of the platform comparing users' trends.

[1] Corresponding Author, Heini Utunen; E-mail: utunenh@who.int.

3. Result

The proportion of women attending the online learning on OpenWHO grew to 51% (from 40%). Still, female learners show lower enrolments (43%) to the health topics other than COVID-19. Users identifying themselves as Other changed from >0.1% to 0.15%. Overall completion rates have increased from 39% before the pandemic, to 54% during the pandemic. The online learning platform has spanned to older and younger user groups. The user group of +70 years now account for 5% of users. Also, the age group of less than 20 years old has grown from 3% pre-pandemic to 11%. OpenWHO's most popular courses, essentially infectious diseases, were highest in the WHO African region (24%). This has decreased to 7%. Examining the platform use based on countries' classification by income level based on the World Bank classification, a remarkable change before and during the pandemic can be witnessed.

Table 1. The platform uses in low-, middle- and high-income countries in 2019 and 2021.

Country classification / % Of enrolments	Low-income countries	Middle-income countries	High-income countries
2019	14.20 %	40.19 %	45.62 %
2021	3.38 %	70.72 %	25.89 %

4. Discussion

More women, younger and older learners used the WHO's learning platform during the pandemic. Public health education leaned toward a more equitable reach including previously underrepresented groups. WHO's platform helped in bypassing the gender inequity in access to education during the pandemic. Our results align with online learning paralleling the rise of the burden of COVID-19. Higher COVID-19 prevalence in high-income countries is coherent with more users in these locations.

5. Conclusions

The pandemic has led to increased learner commitment and new learners from the general public and those vulnerable to the disease, thus expanding and equalizing public health education to previously underrepresented groups. Open online learning ensures the wide access to emergency knowledge for both professionals in public health and to the general public.

References

[1] Dhawan S. Online learning: A panacea in the time of COVID-19 crisis. Journal of Educational Technology Systems. 2020 Sep;49(1):5-22.

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210839

Network Analysis of COVID-19 Vaccine Misinformation on Social Media

Chad MELTON[a,b,1], Olufunto A. OLUSANYA[a] and Arash SHABAN-NEJAD[a,b,1]
[a] *University of Tennessee Health Science Center - Oak Ridge National Laboratory (UTHSC-ORNL) Center for Biomedical Informatics, Department of Pediatrics, Memphis, TN, USA*
[b] *The Bredesen Center, University of Tennessee, Knoxville, TN, USA*

Abstract. Almost half of the world population has received at least one dose of vaccine against the COVID-19 virus. However, vaccine hesitancy amongst certain populations is driving new waves of infections at alarming rates. The popularity of online social media platforms attracts supporters of the anti-vaccination movement who spread misinformation about vaccine safety and effectiveness. We conducted a semantic network analysis to explore and analyze COVID-19 vaccine misinformation on the Reddit social media platform.

Keywords. Vaccines, vaccine hesitancy, social media, misinformation, semantic analysis

1. Introduction

Though social distancing, mask usage, and lockdowns help slow the spread of the COVID-19 virus, there is an overwhelming consensus that vaccinations save lives and remain the best defense against new COVID-19 infections. A plague in itself, the spread of vaccine misinformation and disinformation ultimately lead to unquantifiable negative outcomes (e.g., low vaccination rates, increased infectivity, and hospitalization rates, as well as morbidity, and mortality) from vaccine-preventable diseases [1, 2].

2. Methods

We conducted a semantic network analysis [3] on 5,106 posts displayed by 3,770 unique authors within the Reddit community, *r/NoNewNormal* from November 23, 2020, to July 27, 2021. Our graph was created based on the calculation of *Eigen centrality (EC)* between nodes. EC quantifies and measures the specific influence of a node based on the number of edges it shares with other nodes (i.e., influence) [4].

[1] Corresponding Authors, Chad Melton (Email: chadmeltone@gmail.com) and Arash Shaban-Nejad (E-mail: ashabann@uthsc.edu), Centre for Biomedical Informatics, 50 N. Dunlap St., Memphis, TN 38103, USA.

3. Results

The EC network rendered with its central hub as being the word *vaccine* and reported a centrality value of approximately 0.517. A secondary hub, *people*, was located relatively close to the main hub. This was somewhat expected due to the querying process while harvesting the data. EC values in our graph ranged from [0.002 *survival*, 0.517 *vaccine*]. The distribution of EC values appeared to be Gaussian with a group of four outliers ranging from [0.434, 0.517]. Some subnetworks revealed connections between nodes that could potentially indicate misinformation or hesitancy. The node *effects* were connected to several other nodes including *long, term, vaccine, people, covid, and vaccines.* Upon inspection of our data set, we observed clear detection of a vaccine-hesitant comment related to this network ("I absolutely will not take it and my wife won't either").

4. Discussion and Conclusion

As Proof of Concept, our results suggest that semantic network analysis could be implemented in detecting vaccine misinformation or vaccine hesitancy. Because the causes of vaccine hesitancy are multifaceted, digital intervention technologies should be adopted to surveil and recognize the type of hesitant behavior so an appropriate response can be implemented. These results accompany ongoing research focused on detecting and categorizing vaccine misinformation in social media and other public forums [5]. The information gained from network analysis could be especially useful in designing custom digital technologies [6, 7] to educate vaccine-hesitant users or combat outright dangerous misinformation designed to dissuade users from being vaccinated.

References

[1] Zhang J, Featherstone JD, Calabrese C, and Wojcieszak M. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. Preventive Medicine. 2021;145:106408.

[2] van der Linden S, Sander, Graham Dixon, Chris Clarke, and John Cook. Inoculating against COVID-19 vaccine misinformation. EClinicalMedicine 2021;33:100772.

[3] Shin EK, and Shaban-Nejad A. Applied Network Science for Relational Chronic Disease Surveillance. Stud Health Technol Inform. 2019;262:336-339. doi: 10.3233/SHTI190087.

[4] Developers, NetworkX. "NetworkX documentation." Release 1.7, (2012). Available online at: https://networkx.org/documentation/networkx-1.7/index.htmlRelease

[5] Melton, CA, Olusanya OA., Ammar N, and Shaban-Nejad A. Public Sentiment and Topic Modeling Regarding COVID-19 Vaccines on Reddit Social Media Platform: A Call to Action For Strengthening Vaccine Confidence. Journal of Infection and Public Health. 2021;S1876-0341(21)00228-8. doi: 10.1016/j.jiph.2021.08.010.

[6] Olusanya OA, Ammar N, Davis RL, Bednarczyk RA, and Shaban-Nejad A. A Digital Personal Health Library for Enabling Precision Health Promotion to Prevent Human Papilloma Virus-Associated Cancers. Front. Digit. Health, 21 July 2021. doi:10.3389/fdgth.2021.683161

[7] Ammar N, Bailey JE, Davis RL, Shaban-Nejad A. The Personal Health Library: A Single Point of Secure Access to Patient Digital Health Information. Stud Health Technol Inform. 2020 Jun 16;270:448-452. doi: 10.3233/SHTI200200.

J. Delgado et al. (Eds.)

doi:10.3233/SHTI210840

An Online Information Tool for Diabetic Retinopathy

George AGRIODIMOS[1], Parisis GALLOS, Sotiris TASOULIS and
Ioannis ANAGNOSTOPOULOS
Department of Computer Science and Biomedical Informatics,
University of Thessaly, Lamia, Greece

Abstract. Regardless of the type of diabetes, patients with diabetes are 25 times more likely to develop vision problems or even blindness than non-diabetics. Diabetic Retinopathy is the most common cause of new cases of blindness in adults. The aim of this paper is to present a pilot online tool to provide information regarding the Diabetic Retinopathy. The tool was developed using a Content Management System. To compile the content of the website, a literature review was conducted. The online information tool is addressed to all potential stakeholders on this subject, for the provision of knowledge and targeted information according to their information needs. The online tool also aims to raise the public awareness about the Diabetic Retinopathy and health promotion.

Keywords. Diabetes Mellitus, Diabetic Retinopathy, Content Management System, WordPress

1. Introduction

Diabetes Mellitus (DM) consists of multiple metabolic diseases, which are characterized by chronic hyperglycemia. DM is a widespread worldwide disease and its prevalence has been steadily increasing in recent decades. Diabetic Retinopathy is a serious complication of DM, which occurs damage to the retina due to long term problems of the retinal vessels. Diabetic Retinopathy is the most common cause of new cases of blindness in adults aged 20 to 74 years [1]. Regardless of the type of diabetes, diabetics are 25 times more likely to develop vision problems or even blindness than non-diabetics [2]. Meanwhile, online platforms and Health Informatics seems to play an important role in the management of Ophthalmological Disorders by developing useful tools for prevention and treatment [3,4]. The aim of this paper is to present a pilot online tool to provide information regarding the Diabetic Retinopathy.

2. Methods

In order to develop the proposed tool, a literature review about Diabetic Retinopathy was conducted in relevant printed literature such as medical books and scientific journals as well as in online scientific databases (PubMed, Google Scholar) using

[1] Corresponding Author, George Agriodimos, Student at the Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece; E-mail: giorgosagriodimos@hotmail.com.

relative keywords. Multimedia content was collected from various related websites. WordPress was selected for the tool development and specific plug-ins were installed to improve the functionality of the website's environment, as well as to support the different types of the published content. To cover the users' needs regular reviews of the website development were took place by two experts in the fields of Health Informatics and Informatics. The potential stakeholders and users of this system are healthcare professionals, students, patients, the relatives of them or any other persons who are not aware of the disease and they willing to increase their knowledge. Healthcare professionals can add content related to the Diabetic Retinopathy on the website. Students, patients and relatives can make comments on the published content. To ensure the content reliability, the uploaded content and the users' comments are under review before publishing.

3. Results and Discussion

A pilot online tool to inform about Diabetic Retinopathy was developed in Greek language. The users can access this tool easily using a simple web browser. The content of this information platform includes images, text, videos, figures regarding the Diabetic Retinopathy, as well as links to recent scientific publications in this domain. This content is fed from the administrator or other super user-experts. In addition, visitors can be registered to this system to be able to post articles and comments, as well as, to edit the content based on their role. The usage of this tool can lead to the formulation of an online community of people who are interested in Diabetic Retinal (Eye) Disease including also patients. The tool offers a digital "place" where people of different disciplines can exchange their views and opinions in matters about the disease. The users can also update the content of the website after the administrator's or author's consent.

4. Conclusions

The presented online pilot tool was created in order to inform the public about a particular condition that is not widely known and to increase the awareness for Diabetic Retinopathy. Strong knowledge about the disease can lead to early detection and better outcome for the patients. This pilot tool was developed in the context of the bachelor thesis. Future work includes the further development of the tool, the public access of it, as well as, the dissemination of the tool among potential users and various disease-related bodies such as scientific companies or patient associations.

References

[1] Kanski JJ, Bowling B. Kanski's clinical ophthalmology e-book: a systematic approach. Elsevier Health Sciences. 2015.

[2] Barrett EJ, et al. Diabetic Microvascular Disease: An Endocrine Society Scientific Statement. J Clin Endocrinol Metab. 2017 Dec 1;102(12):4343-4410.

[3] Lin SH, Lin TM. Demand for online platforms for medical word-of-mouth. Journal of International Medical Research. 2018; 46(5):1910-1918.

[4] Patte M, Liaskos J, Gallos P, Mantas J. An Online Tool for Ophthalmological Disorders. Studies in health technology and informatics 2020; 270:1197-1198.

J. Delgado et al. (Eds.)

Subject Index

Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

Author Index