German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210068

Preliminary Analysis of Structured Reporting in the HiGHmed Use Case Cardiology: Challenges and Measures

Aljoscha KINDERMANN[a,b,1], Erik TUTE[c], Sebastian BENDA [b], Martin LÖPPRICH[d], Phillip RICHTER-PECHANSKI [a,b] and Christoph DIETERICH[a,b,1]
[a] *Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg*
[b] *Department of Internal Medicine III, University Hospital Heidelberg*
[c] *Peter L. Reichertz Institute for Medical Informatics of the University of Braunschweig - Institute of Technology and Hannover Medical School, Germany*
[d] *Center for Information Technology and Biomedical Engineering, Heidelberg University Hospital, Heidelberg, Germany*

Abstract. The HiGHmed consortium aims to create a shared information governance framework to integrate clinical routine data. One challenge is the replacement of unstructured reporting (e.g. doctoral letters) with structured reporting in clinical routine. The Heidelberg cardiology department evaluates dynamic PDF forms for structured data reporting of heart failure (HF) patients. In this use case, we aim to identify potential caveats or shortcomings in data processing at an early stage. We employed data mining strategies to detect patterns related to incomplete or false data, which we found to be present among all data types. We then discuss the characteristics of the baseline patient cohort in Heidelberg to find out about specific peculiarities and potential biases, which may be site-specific. Briefly, our patient population is predominantly male (67%), NYHA I & II are the most common severity classes, NYHA IV is missing entirely. Most patients have a dilated cardiomyopathy (DCM) or coronary heart disease (CHD) diagnosed as their cause of HF. Finally, we also analyzed how comorbidities and risk factors relate to specific disease entities of heart failure patients. Family anamnesis was more frequent among cardiomyopathy patients than among CHD patients, who show a more dominating presence of dyslipidemia instead. Generally, the most dominant risk factor was arterial hypertension, while at the other end of the scale alcoholism appears to be underreported.

Keywords. HiGHmed, heart failure, electronic health records, openEHR, data quality, missing data, measurement methods, data analysis, risk factor, comorbidity

1. Introduction

The potential of reusing clinical data for research is fueled by the growing adoption of electronic health records (EHR) [1]. These data, however, are often stored in distributed and heterogeneous systems. The HiGHmed consortium aims to unite these data contents

[1] Corresponding authors: AK and CD, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany aljoscha.kindermann@med.uni-heidelberg.de, christoph.dieterich@med.uni-heidelberg.de.

into medical data integration centers (MeDICs) to make them available for collaborative use by participating partner sites [2]. To achieve interoperability, data are stored compliant to the openEHR specification, structured according to templates which address a respective clinical situation and also provide constraints on the data, e.g. datatypes, eligible values, cardinalities etc. [3].

The HiGHmed Use Case Cardiology aims to get a better understanding of the multifaceted disease characteristic of heart failure (HF) by improving the handling of data from clinical routine. Among other data sources, we extract information from discharge letters, which are unstructured documents. Discharge letters contain most valuable information, such as physicians' patient assessment. For this purpose, we explore alternative documentation approaches using electronic structured forms. Specifically, the Heidelberg cardiology department implemented dynamic PDFs (dPDFs) into the hospital information system. These electronic forms offer a human readable user interface as well as a structured XML backend which is transmitted via an ETL process to the Medical Data Integration Centre (MeDIC). At the time of this analysis, 168 patients were already recorded using dPDF forms. We use this initial data set to evaluate the recoding process and the recorded data to identify potential shortcomings or caveats as it is recommended for distributed research networks in the literature [4]. In the course of our analysis, we also assessed data patterns related to data missingness, compliance with the openEHR template requirements and how general patient characteristics can uncover potential biases or shortcomings. Finally, we present the distribution of patient subgroups (i.e. type of heart failure) as well as presence and co-occurrence of additional risk factors and compare our findings to literature reports.

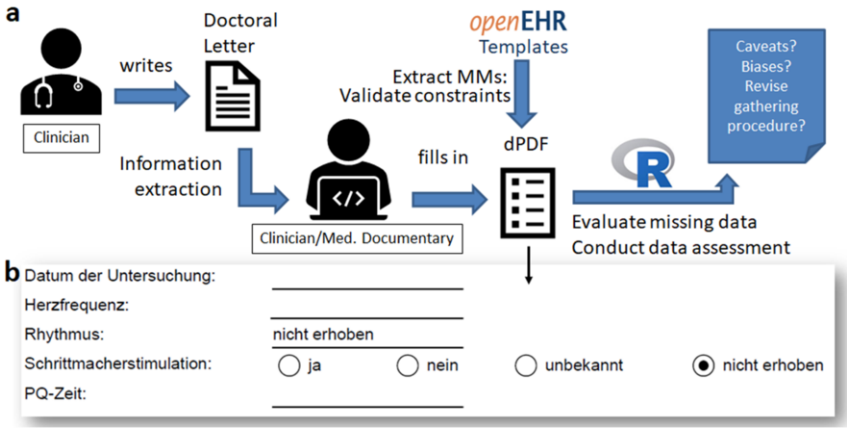

Figure 1: Schematic representation of the current dPDF documentation process (a) and a section of the displayed dPDF document (b).

2. Methods

2.1. Dynamic PDF – Structure and Recording Process

The dynamic PDF (dPDF) provides an electronic form which can be divided into the following sections: A Header, storing patient name and other information which will remain unchanged over time. Anamnesis data, which the doctor will enquire and measure

on the patient's visit. Further sections are relevant medication classes as well as selected MRI, ECG and cardiac catheter fields that cannot be retrieved from other databases. Finally, we also capture a selection of end-point variables such as death of the patient and severe cardiac events.

Within these sections data can be recorded using different data types: Date fields, numeric fields, free text fields, dropdown selection fields and boolean fields, which offer "yes" and "no" as possible options, but also "unknown" and "not acquired" for missing data (Figure 1b, suppl. Figure 1)[2]. The data which is filled in the dPDF is stored as human readable PDF and also as XML structure.

The dPDF was designed in such a way that a physician would fill in the data in production mode. Our prototype was filled in by a domain expert who obtained the data primarily from doctoral letters of the participating patients (see Figure 1a).

2.2. Data extraction and analysis

We use R version 3.6.2 for data extraction, preprocessing and analysis. We use openEHR templates, which were collaboratively defined by all partner sites of the HiGHmed use case cardiology. These templates contain formalized constraints for valid variable values that we check in the dataset. Applying the tool openCQA, we automatically derive measurement methods (MMs) which are R-scripts assessing quality measures based on formalized knowledge [5][3], e.g. for checking these constraints or providing some descriptive measures like boxplots. These MMs are used to evaluate the dPDF data in an initial quality control and curation to enforce the same data standards across the consortium. Since openCQA originally aims at data quality assessment for openEHR-based data repositories, its utilization for this preliminary data analysis on data in XML structures generated from the dPDFs also tests the transfer of the openCQA-MMs to comparable data in other representations.

3. Results

3.1. Missing data patterns

One cornerstone of our analysis was to understand which fields of the dPDF were often not filled in (i.e. missing). Patterns of not randomly missing data can indicate structures in the data gathering process that need to be improved.

Figure 2a depicts the percentage of missing values in date/time, free text and numeric fields. The plot shows some clear patterns, which differ between the data types. Among these, we found fields that were of special importance due to their content or missingness pattern. With 36 % NA, the field "Other medication" stands out in contrast to most other free text fields, as they are usually filled for all patients (e.g. "Patient Name"). Numeric fields seem to split into three main visual clusters, where some are almost never empty (among these for example the "ECG QT time"). In the second cluster we find "MRI cardiac output", with 67 % NA. Interestingly, by contrast, "MRI T1 time",

[2,3] Supplement: dPDF examples and openCQA repository, respectively:
https://github.com/dieterich-lab/Analysis-of-structured-reporting/blob/master/README.md
https://gitlab.plri.de/tute/openehr-dq

derived from the same measurement, appears to belong to another visual cluster and is empty for 99 % of the patients.

Not depicted in that plot are the boolean fields and entries that can be selected by dropdown in the form, which by definition can never be empty. However, these fields can be set by the user to "unknown" or "not acquired" which is equally defining a missing value. The percentage of times that each boolean and dropdown field has been set to either one of these missingness options is shown in Figure 2b. Among the fields set to "not acquired", we find of particular interest that "diagnosed alcoholism" has a high number of patients with missing data (79 % NA) as well as "depression" (67 % NA). The interpreted MRI field "MRI LGE suggests infarction" could not be acquired in 64 % of the cases. By contrast, fields with high coverage are information of "Smoking status" (8 % NA) and "Pacemaker Stimulation" (10 % NA).

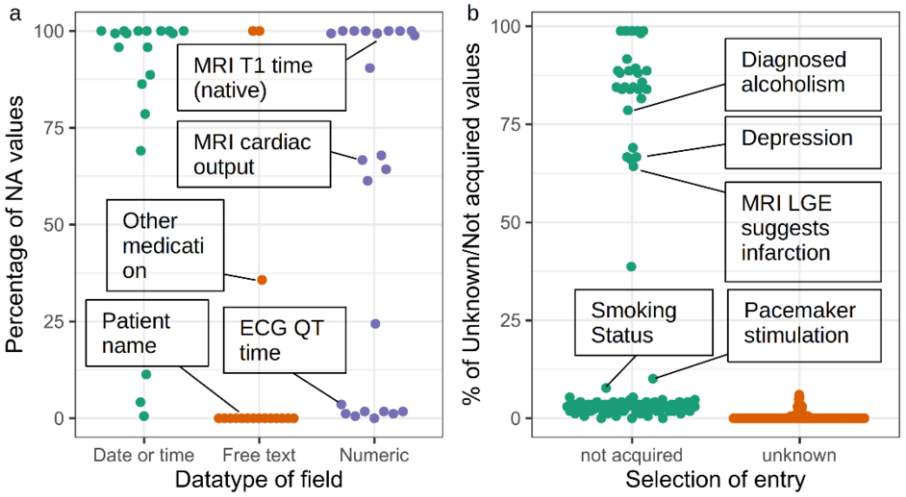

Figure 2. Each dot represents one data field of the dPDF. a) Percentage of NA values per field b) Percentage of value type "not acquired" or "unknown" for each boolean or selection field.

3.2. Measurement methods derived for common templates

MMs were derived for a cardiologic anamnesis template resulting in over 1000 R-scripts representing potential MMs. Out of these we selected 15 R-scripts which were reasonably applicable to the numeric anamnesis variables in the dPDF. The selected analyses comprised boxplots and minimum, maximum, mean, median, modus and quantiles for these variables. Three MMs explicitly targeted particular variables checking collaboratively defined constraints in the templates. All dataset values matched these constraints.

3.3. Data content assessment – Basic patient characteristics

In addition to analyzing data quality and missing data patterns in the dPDF recorded data, it was also necessary to study the recorded patient and disease characteristics for specific patterns and potential biases.

Table 1 shows the basic characteristics of the enrolled patients. Two thirds of the cohort were of male sex. The age group between 60 and 79 is with 42 % most abundant

in our data (Average 54.6 years and 59.3 years, female, male, respectively). NYHA II was the most abundant class (41%), however almost equal to NYHA I (40%). NYHA IV was not present within our patient population.
The dominating cardiac disease present among the patients was coronary heart disease (CHD). 53 % of patients showed this disease, however, it was only diagnosed as the primary cause of their HF symptomatology for 22 % of patients. The second most common primary HF cause was dilated cardiomyopathy (DCM), followed by hypertrophic cardiomyopathy (HCM). Other causes add up to a total of 24 %.

Table 1. Characteristics of the patient population and their respective subgroups

Characteristics	**n = 168**	**Total Sample**
Subgroup	**n**	**%**
Gender		
Men	113	(67)
Women	55	(33)
Age		
20-39	28	(17)
40-59	55	(33)
60-79	70	(42)
80-100	15	(9)
NYHA Class		
I	67	(40)
II	69	(41)
III	26	(15)
IV	0	(0)
HF relevant classification of Disease		
DCM	70	(42)
HCM	20	(12)
NCCM	5	(3)
ARVC	2	(1)
CHD	37	(22)
Others	34	(20)
Cardiovascular Risk Factors		
Arterial Hypertension	87	(52)
Dyslipidemia	64	(38)
Obesity	44	(26)
Current Smoker	18	(11)
Former Smoker	55	(33)
Alcohol Use Disorder	2	(1)
Kidney Disease	24	(14)
Diabetes	33	(20)
Family Medical History	64	(38)
Former Cancer Patient	20	(12)
Peripheral Artery Disease	5	(3)

3.4. Risk factors & their relation to disease characteristics

The dPDF documents the most important cardiovascular risk factors. As shown in Table 1, Arterial Hypertension was the most abundant risk factor (52 %) among the enrolled HF patients, followed by dyslipidemia and a positive family anamnesis with both 38 %. Alcohol Use Disorder (1%) and Peripheral Artery Disease (3%) were rarely recorded.

In addition to the mere distribution of patient subgroups we were interested in possible relationships between diseases and the reported risk factors among the patients. Figure 3a shows that the number of risk factors per patient increased with the NYHA

class level. Figure 3b-d indicates how many patients had their respective HF causing disease co-occurring with risk factors and co-morbidities. HCM patients very often also showed a positive family anamnesis as well as hypertension, or obesity and dyslipidemia as major risk factors. These prominent patterns were also found among DCM patients. CHD patients showed a frequent "triangular" co-occurrence of hypertension and dyslipidemia. CHD and obesity were a less common combination among our patients, also family anamnesis was a less frequent risk factor compared to DCM. Valve diseases, myocardial infarctions and atrial fibrillation appeared all to be frequent comorbidities among CHD patients.

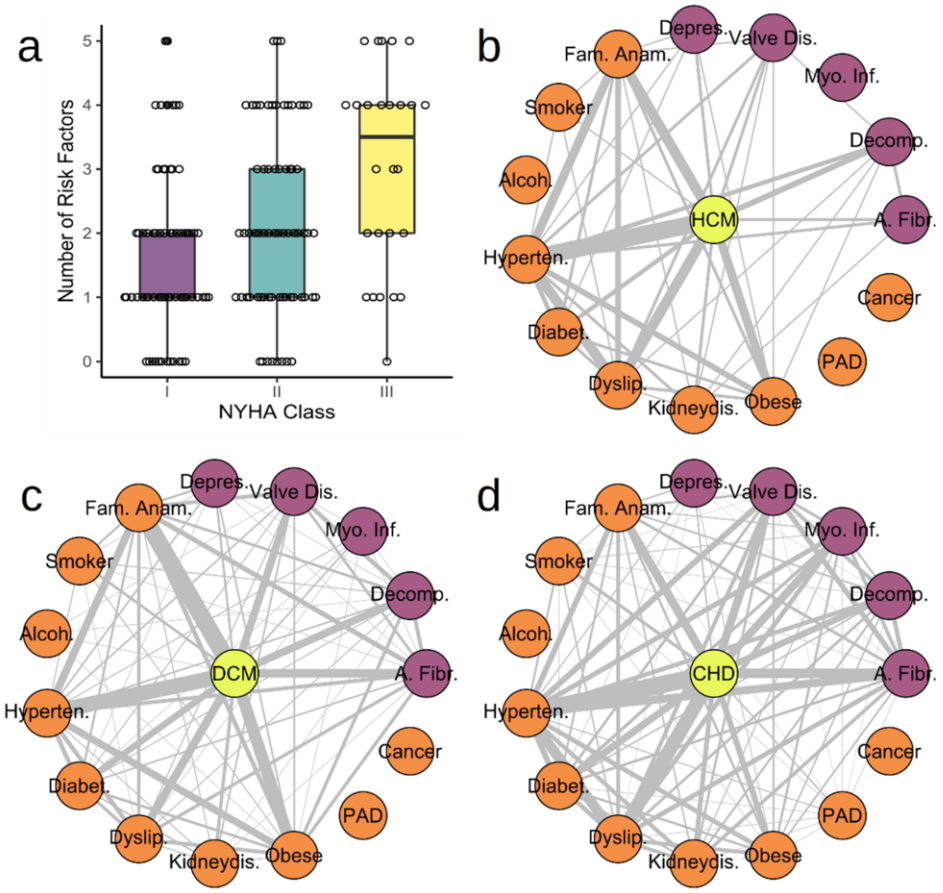

Figure 3. a: The number of positively noted risk factors per patient increases from NYHA I to III. b-d: In the network representation, the edge thickness between the nodes indicates the number of patients who have DCM, HCM or CHD (center node) co-occurring with the shown co-morbidities and risk factors (surrounding nodes). (Abbreviations: Depres.: Depression, Valve Dis.: Valve disease, Myo. Inf.: Myocardial infarction, Decomp.: Cardiac decompensation, A. Fibr.: Atrial fibrillation, PAD: Peripheral artery disease,, Kidneydis.: Kidney disease, Dyslip.: Dyslipidemia, Diabet.: Diabetes mellitus, Hyperten.: Hypertension, Alcoh.: Diagnosed alcoholism, Fam. Anam.:Positive family anamnesis, DCM: Dilated cardiomyopathy, HCM: Hypertrophic cardiomyopathy, CHD: Coronary heart disease)

4. Discussion

4.1. Putting patient characteristics into context

The patient cohort of the first HiGHmed patients consists to 2/3 of male patients. An overabundance of men in HF patients is consistent with what is known about HF cohorts ([6], [7]). Interestingly, female HF patients are usually older than their male counterparts on average, which was not the case for our patients. Another difference to known patterns from literature is the high frequency of NYHA I among patients[8]. On the other hand, the risk factors that were most abundant among the dPDF patients align well with the risk factors that have been found most relevant for German heart patients in the past [9]. Hypertension, often reported as the number one risk for heart problems, is also the most prevalent among our cohort. Interestingly, the percentage among our patients matches almost exactly the prevalence reported in the GEMCAS study [10]. Furthermore, the combination with dyslipidemia is a common pattern among CHD patients. The inclusion of more patients in the future might rule out some of the deviations found above.

Through the usage of openCQA, we verified and confirmed that the dPDF data complies with the openEHR template constraints defined by the consortium. Further, we could show that openCQA-MMs, which originally work on openEHR data repositories, can be transferred with little effort to other data representations.

4.2. Sources of missing data explained

Patterns we found indicate potential for improvements in the data gathering process. The low number of patients suffering from alcohol use disorder is surprising, given a suggested prevalence between 23 % and 47 % in DCM (cf. Rehm et al. [11]). The author also describes the phenomena of underreporting which is likely the explanation among our patient cohort. This might just as well be the explanation for the high number of NA values for the comorbidity depression. In both cases underreporting can potentially be explained by the stigma associated with the diseases. The compromise between documenting risk factors and maintaining patient discretion is not easy to find. In other cases of missing data, the reasons are more straight forward. We found a complete absence of NYHA IV patients, which is likely due to the recruiting procedure within the outpatient cardiology department, while NYHA IV patients are mostly bedbound and therefore usually inpatients. Likewise, we also see a low number of patients of age greater than 80. Older and more severely impaired patients might be more hesitant or incapable to participate in clinical studies. We will have to find ways not to exclude specific subgroups of the HF population during data gathering.

Other sources of missing data can be identified for the MRI data, where some fields had missing values for 64 % - 67 % of the patients, which can be explained with the lower clinical indication of MRI examinations. The fact that not every examination is necessary for every patient is an explanation for missing data among some of the dPDF data sources. Interestingly, we also found MRI fields with 99 % NA. Reconciliation with our clinical documentation revealed, that these fields were rarely transferred from the database to the discharge letter. Hence, a major challenge will be to link in primary databases to dPDF records. The manual transfer of data from the primary sources to the forms by a documentary is another potential error source that should be investigated in future analyses.

5. Conclusion

Our analysis showed that our preliminary data collected with the dPDF agrees with trends found in literature about HF patient cohorts. Our data also suggest that some patient subgroups are underrepresented and should be included in the recruiting process in the future. Likewise, some scarce data sources should be proactively obtained from other (primary) data sources to the dPDF. Underreporting of risk factors such as alcoholism is difficult to prevent, however might be tackled by an adjusted patient enquiry. Overall, our analysis gives a glimpse into the possibilities of data reuse when captured in a structured format. Ongoing research in our group focusses on the possibilities of creating semi-automated doctoral letter texts from dPDF content.

6. Conflict of Interest: None

7. Acknowledgement

We would like to thank all members of the Dieterich Lab and the Use Case Cardiology team for their great input and insightful discussions. The work of A.K., P.R.P. and C.D. was kindly supported by the BMBF-funded HiGHmed consortium (Medical Informatics Initiative Germany). C.D. would like to thank the Klaus Tschira Stiftung gGmbH (grant 347 00.219.2013) for providing all necessary compute infrastructure.

References

[1] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C. U. Lehmann, "Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress," *Yearbook of medical informatics*. 2017.

[2] B. Haarbrandt *et al.*, "HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries," *Methods Inf. Med.*, vol. 57, no. S 01, pp. e66–e81, Jul. 2018.

[3] J. L. Cardoso de Moraes, W. L. de Souza, L. F. Pires, and A. F. do Prado, "A methodology based on openEHR archetypes and software agents for developing e-health applications reusing legacy systems," *Comput. Methods Programs Biomed.*, vol. 134, pp. 267–287, Oct. 2016.

[4] G. Welch, F. Von Recklinghausen, A. Taenzer, L. Savitz, and L. Weiss, "Data Cleaning in the Evaluation of a Multi-Site Intervention Project," *eGEMs*, 2017.

[5] E. Tute and M. Marschollek, "Technical architecture of a tool for interoperable data characterization," in *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS); 2019 September 8; Dortmund, Germany*, 2019.

[6] C. Magnussen *et al.*, "Sex-Specific Epidemiology of Heart Failure Risk and Mortality in Europe: Results From the BiomarCaRE Consortium," *JACC Hear. Fail.*, vol. 7, no. 3, pp. 204–213, 2019.

[7] P. A. Mehta and M. R. Cowie, "Gender and heart failure: A population perspective," *Heart*, vol. 92, no. SUPPL. 3, pp. 14–18, 2006.

[8] A. Ahmed, W. S. Aronow, and J. L. Fleg, "Higher New York Heart Association classes and increased mortality and hospitalization in patients with heart failure and preserved left ventricular function," *Am. Heart J.*, 2006.

[9] Deutsche Herzstiftung e.V., "28. Deutscher Herzbericht - Sektorenübergreifende Versorgungsanalyse zur Kardiologie, Herzchirurgie und Kinderherzmedizin in Deutschland," 2016.

[10] C. Balijepalli, P. Bramlage, C. Lösch, C. Zemmrich, K. H. Humphries, and S. Moebus, "Prevalence and control of high blood pressure in primary care-results from the German metabolic and cardiovascular risk study (GEMCAS)," *Hypertens. Res.*, 2014.

[11] J. Rehm, O. S. M. Hasan, S. Imtiaz, and M. Neufeld, "Quantifying the contribution of alcohol to cardiomyopathy: A systematic review," *Alcohol*, vol. 61, pp. 9–15, 2017.