Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220507

A Federated and Distributed Data Management Infrastructure to Enable Public Health Surveillance from Intensive Care Unit Data

Jonas BIENZEISLER[a,1], Lucas TRIEFENBACH[a], Alexander KOMBEIZ[a], Matthäus LOTTES[b], Christopher VOGEL[b], Linus GRABENHENRICH[b], Martina FISCHER[b], Theresa KOCHER[b], Lukas NIEKRENZ[c], Michael DREHER[c], Christoph MÜLLER[d], Rainer RÖHRIG[a] and Raphael W. MAJEED[a] on behalf of the AKTIN and SPoCK Research Group

[a]*Institute of Medical Informatics, Medical Faculty of RWTH Aachen, Aachen, Germany*
[b]*Department of Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany*
[c]*Department of Pulmonology and Intensive Care Medicine, Medical Faculty of RWTH Aachen, Aachen, Germany*
[d]*Data Integration Centre (DIC), University Hospital RWTH Aachen, Aachen, Germany*

Abstract. The Robert Koch Institute (RKI) monitors the actual number of COVID-19 patients requiring intensive care from aggregated data reported by hospitals in Germany. So far, there is no infrastructure to make use of individual patient-level data from intensive care units for public health surveillance. Adopting concepts and components of the already established AKTIN Emergency Department Data registry, we implemented the prototype of a federated and distributed research infrastructure giving the RKI access to patient-level intensive care data.

Keywords. Critical care, routine care research, EHR, database, data privacy

1. Introduction

Intensive Care Units (ICU) are not only a crucial part of the health system but also an essential data source for public health surveillance. For the modeling and prediction of ICU occupancy during the COVID-19 pandemic, German authorities rely on mandatory reporting of COVID-19 cases and aggregated capacity data from hospitals [1]. Such data are collected and analyzed by the Robert-Koch Institute (RKI), the German National Public Health Institute. In the Project SPoCK (*Steuerungs-Prognose von intensivmedizinischen COVID-19-Kapazitäten*), the RKI, cooperating with universities, aims at improving the prediction of intensive care COVID-19 capacities. The automated provision of routine data sources on an individual-patient level is expected to enhance the predictive precision of the models. Furthermore, such data is seen of great value to

[1] Jonas Bienzeisler, Corresponding author; E-mail: jbienzeisler@ukaachen.de

improve *public health surveillance* in general [2]. Internationally, routine data could also be re-used for comparative effectiveness research or evaluation of COVID-19 government interventions. However, assessment of data on individual-patient level from multiple sites is opposed by legal guidelines in the European Union. The General Data Protection Regulation (GDPR) limits the sharing of sensitive data, typically requiring explicit patient consent for data processing. Our objective was to develop and operate a GDPR-compliant research infrastructure for routine data stemming from ICUs, that enables public health surveillance and other applications through the RKI.

1.1. Background

In Germany, data from the intensive care registry[2] is used to track the number of COVID-19 patients requiring intensive care [1]. Aside from structural characteristics of hospitals that maintain intensive care beds, information on ICU capacities and COVID-19 case numbers are collected in a *central* database. For reimbursement, it is also mandatory for ICUs to collect vital parameters and multiple clinical scores in patient data management systems. In practice, clinical scores are used to assess the current, disease-related patient status for individual risk prediction and necessary resources [3]. The predictive value of these scores has also been demonstrated in patients with COVID-19 admitted to the ICU [4] and could be used in occupancy models for the management of ICU capacities [2].

These data cannot be used so far because centralized storage of individual health data under the GDPR typically requires consent which is not feasible for ICU patients and might cause a selection bias. Privacy-preserving methods are usually used to access and analyze sensitive medical data for research despite data protection regulation. Two strategies are common to make use of data; either algorithms are distributed to the decentralized data and only anonymous results are aggregated [5,6] or only data that can be considered anonymous is aggregated (as in the intensive care registry) [7,8].

The *AKTIN* (*Alliance for information and communication technology in intensive* care *and emergency medicine*) *Emergency Department Data Registry* demonstrates, that such an evaluation is possible using a federated and distributed research infrastructure based on *decentral* data warehouses for anonymous aggregation or distributed computing [7–9]. Employing the concepts and software solutions of the AKTIN Registry, we implemented a data management infrastructure for the evaluation of ICU routine records for health surveillance, the so-called *SPoCK Data-Infrastructure*.

2. Methods

The methods for concepting the SPoCK Data-Infrastructure are based on the specific requirements for data capture and general requirements for sustainable operation.

2.1. Specific Requirements for data capture

The RKI required fast and actual access to patient-level data from multiple German ICUs for public health surveillance of COVID-19. The central analysis of structured reports

[2] www.intensivregister.de

by the RKI was mandatory, as the data needs to be analyzed together with pre-existing surveillance models. Thus, standardized processes for retrieval, transmission, and analysis of collected data were necessary. Semantically and syntactically comparable and compatible data were required. Finally, participating ICUs needed the tooling and implementation guidelines to store data and render it accessible for central pooling.

2.2. General Requirements

The handling of the health data is subject to legal and ethical constraints varying from country to country. In the European Union and under the GDPR, consent is mandatory for processing health data, if opening clauses are not in place. Under the premise of proper technical and organizational measures, health data may be processed for reasons of public interest such as health surveillance (Art. 9 (2) (i), GDPR). Integration into national and international research efforts is mandated for sustainable infrastructure projects. The generic concepts of the telematics platform for medical research networks [10] and the German medical informatics initiative [11] determine a national state of the art. Common technologies, terminologies, and data models are obligatory for integration into international research efforts.

3. Results

The general and specific requirements of the data collection match the requirements of the AKTIN Registry [9]. For setting up the SPoCK Data-Infrastructure in a short time, components of the AKTIN Registry were therefore modified to suit the use case. As there were only minor resources for data collection in participating hospitals, we restricted data to routine records collected for billing purposes. We aligned our concepts with recommendations for data protection for medical research networks in Germany [10].

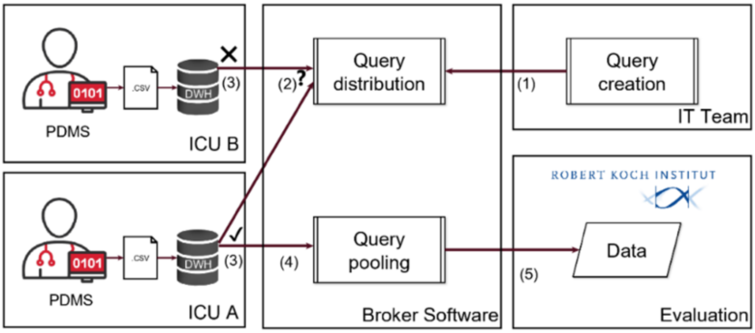

Figure 1: Federated and distributed research infrastructure of the Project. (1) The IT Team creates a query. (2) The query is distributed by the broker software to participating ICUs and can be checked (3) locally. (4) The SPoCK data warehouse (DWH) submits the results of the query to the broker software, anonymized data exports are pooled. (5) Data are delivered to the RKI.

3.1. Concept

We used the pre-existing software solutions of the AKTIN infrastructure for distributed data collection and federated data storage (c.f. Fig. 1). The used components are *not* connected to the AKTIN Registry but form a separate infrastructure. Data are stored in modified instances of AKTIN data warehouses that employ an individual *SPoCK import script* [7,9]. The import script consists of an *Extract-Transform-Load* (ETL) pipeline employing Logical Observation Identifiers Names (LOINC) and Codes and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for annotation in i2b2. A modified instance of the AKTIN Broker can be used to query data. The data warehouse software is supplied and deployed using Linux-based installation packages.

Using these software components, a data set can be collected continuously and prospectively for the purposes of public health surveillance in a data warehouse. For simplicity, required data are exported in a CSV format from the patient data management system daily and then imported into the data warehouse. Each clinic operates and administrates the data warehouse on a dedicated server. The ownership of data and responsibility lies with the respective clinic. For public health surveillance, the data can be queried centrally via the broker and then evaluated by the RKI. It is the obligation of participating clinics to ensure that queried data are anonymous (c.f. Fig. 2).

3.2. Implementation

Currently, the framework is being piloted in the first clinic after approval of the ethics committee[3]. Data from 13144 cases were imported into the data warehouse. The first delivery of data to evaluation is intended for the first quarter of 2022. Furthermore, it is planned to roll out the software to two more clinics operating patient data management systems from different vendors.

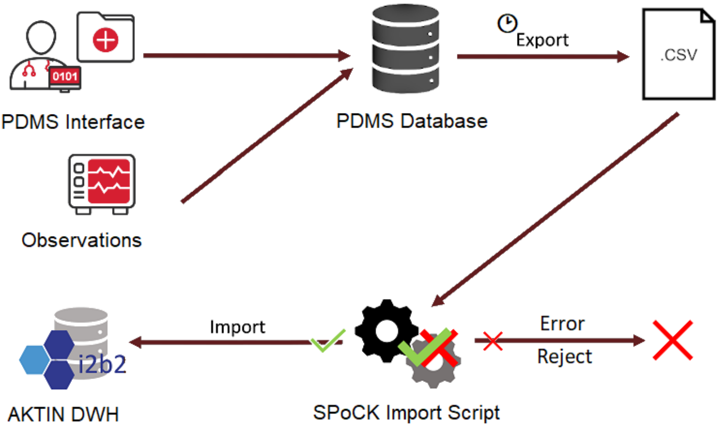

Figure 2: Data collection. Data are exported from patient data management system in CSV format using a repeating Cronjob. Exports are automatically imported into a AKTIN data warehouse (DWH) using the SPoCK import script which consists of an Extract-Transform-Load (ETL) pipeline.

4. Discussion and conclusions

The primary objective of our work was to enable public health surveillance from patient-level ICU data. We designed a research infrastructure that is capable of automatically providing the RKI with such data. We started with data collection and proved feasibility

[3] Ethics Committee Medical Faculty of RWTH Aachen University, EK 483/21.

in practice. The infrastructure is in operation. We based the infrastructure on pre-existing software solutions of the AKTIN Registry. The proof was given that the concepts and software components of the AKTIN Registry can be adapted to other scenarios and are ready for use in scenarios besides the emergency department. Further, the AKTIN components carry the potential for the rapid set up of health surveillance protocols. One advantage is that neither patients' informed consents need to be collected nor data needs to be aggregated on client side. As a result, the infrastructure can be scaled more easily to include data from small and large hospitals nationwide. The only prerequisite is automatic data provision and the installation of an AKTIN data warehouse. As long as the same terminologies and data models are used. it is possible to transfer the approach to other countries; import scripts or interfaces must be adapted accordingly.

Due to financial, organizational, and technical limitations, our work is limited to data collected from one ICU. Multiple instances have yet to be connected to a research network to provide data from multiple sites. Currently, data only consists of clinical scores collected for billing purposes. In the future, the data set needs to be extended and standardized to guarantee syntactic and semantic interoperability. Industry standards like HL7 ORO messages or HL7 FHIR resources should be used instead of CSV format to allow data stemming from any data source, nationally and internationally. Data are only collected for public health surveillance and cannot be used for health research. However, the infrastructure could be adapted to allow for potential health research as well.

References

[1] Schuppert A, Theisen S, Fränkel P, Weber-Carstens S, Karagiannidis C. Bundesweites Belastungsmodell für Intensivstationen durch COVID-19. Med Klin Intensivmed Notfmed 2021.

[2] Azzopardi-Muscat N, Kluge HHP, Asma S, Novillo-Ortiz D. A call to strengthen data in response to COVID-19 and beyond. J Am Med Inform Assoc 2021;28:638–39.

[3] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian J Crit Care 2014;18:220–28.

[4] Taylor EH, Marson EJ, Elhadi M, Macleod KDM, Yu YC, Davids R, et al. Factors associated with mortality in patients with COVID-19 admitted to intensive care: a systematic review and meta-analysis. Anaesthesia 2021;76:1224–32.

[5] 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 Intellegence 2020;2:96–107.

[6] Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced Rate of Inpatient Hospital Admissions in 18 German University Hospitals During the COVID-19 Lockdown. Front Public Health 2020;8:594117.

[7] Brammen D, Greiner F, Kulla M, Otto R, Schirrmeister W, Thun S, et al. The German Emergency Department Data Registry – real-time data from emergency medicine: Implementation and first results from 15 emergency departments with focus on Federal Joint Committee's guidelines on acuity assessment. Med Klin Intensivmed Notfmed 2020.

[8] Boender TS, Cai W, Schranz M, Kocher T, Wagner B, Ullrich A, et al. Using routine emergency department data for syndromic surveillance of acute respiratory illness before and during the COVID-19 pandemic in Germany, week 10-2017 and 10-2021, Preprint at medrxiv 2021.

[9] Ahlbrandt J, Brammen D, Majeed RW, Lefering R, Semler SC, Thun S, et al. Balancing the need for big data and patient data privacy--an IT infrastructure for a decentralized emergency care research database. Stud Health Technol Inform 2014;205:750–54.

[10] Pommerening K., Helbing K, Ganslandt T, Drepper J Leitfaden zum Datenschutz in medizinischen Forschungsprojekten: Generische Lösungen der TMF 2.0. Schriftenreihe der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V, Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG; 2017.

[11] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research. Methods Inf Med 2018;57:e50-6.