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


*doi:10.3233/SHTI210075*

# Needs for an Integration of Specific Data Sources and Items – First Insights of a National Survey Within the German Center for Infection Research

Carolin E.M. JAKOB[a,b], Melanie STECHER[a,b], Sandra FUHRMANN[a,b], Sebastian WINGEN-HEIMANN[b,c], Stephanie HEINEN[a,b], Gabriele ANTON[a,d], Michael BEHNKE[e], Uta BEHRENDS[a,f], Martin BOEKER[g], Stefanie CASTELL[a,h], Hans DEMSKI[d], Maximilian DIEFENBACH[a,i], Jane C. FALGENHAUER[a,j], Moritz FRITZENWANKER[a,j], Petra GASTMEIER[a,e], Markus GERHARD[a,k], Stephan GLÖCKNER[a,h], Mira GOLUBOVIC[l], Barbara GUNSENHEIMER BARTMEYER[m], Josef INGENERF[n], Rolf KAISER[b], Marie-Luise KÖRNER[h], Wibke LOAG[a], Alice MCHARDY[a,h], Ernst MOLITOR[o], Ulrich NÜBEL[a,h], Michael PRITSCH[i], Michael RAMHARTER[a,p], Sigbert R. RIEG[q], Jan RUPP[a,r], Daniela SCHINDLER[a,k], Dominik SCHWUDKE[s], Christoph SPINNER[a,k], Benjamin STOTTMEIER[a], Maria VEHRESCHILD[a,l], Matthias WILLMANN[a,t], and Jörg J. VEHRESCHILD[a,b,l,1]

[a]*German Center for Infection Research (DZIF), Braunschweig*
[b]*Department I of Internal Medicine, University of Cologne, Cologne*
[c]*FOM, University of Applied Sciences, Cologne*
[d]*Institute of Epidemiology, Helmholtz Zentrum Muenchen, Munich*
[e]*Inst. of Hygiene and Environmental Medicine, Charité Universitätsmedizin Berlin*
[f]*Children's Hospital, Technical University Munich, Munich, Germany*
[g]*Inst. of Med. Biometry & Statistics, University of Freiburg, Freiburg*
[h]*Helmholtz Zentrum for Infection Research, Braunschweig*
[i]*Dev. of Infectious Diseases and Tropical Medicine, University Hospital LMU, Munich*
[j]*Institute of Medical Microbiology, Justus-Liebig-University, Gießen*
[k]*School of Medicine, Technical University of Munich, Munich*
[l]*Department of Internal Medicine, Goethe University Frankfurt, Frankfurt am Main*
[m]*Robert-Koch Institute, Berlin*
[n]*Institute of Medical Informatics, University of Luebeck, Luebeck*
[o]*Inst. for Microbiology, Immunology and Parasitology, University Bonn, Bonn*
[p]*Department of Tropical Medicine, University Medical Center Hamburg-Eppendorf*
[q]*Department of Medicine II, University of Freiburg, Freiburg*
[r]*Department of Infectious Diseases and Microbiology, University of Luebeck, Luebeck*
[s]*Research Center Borstel*
[t]*Institute of Medical Microbiology and Hygiene, University of Tuebingen, Tuebingen*
***All Affiliations in Germany***

[1] Corresponding Author, Prof. Jörg Janne Vehreschild, Department I for Internal Medicine, University Hospital of Cologne, Herderstraße 52-54, Joerg.vehreschild@uk-koeln.de

**Abstract.** State-subsidized programs develop medical data integration centers in Germany. To get infection disease (ID) researchers involved in the process of data sharing, common interests and minimum data requirements were prioritized. In 06/2019 we have initiated the German Infectious Disease Data Exchange (iDEx) project. We have developed and performed an online survey to determine prioritization of requests for data integration and exchange in ID research. The survey was designed with three sub-surveys, including a ranking of 15 data categories and 184 specific data items and a query of available 51 data collecting systems. A total of 84 researchers from 17 fields of ID research participated in the survey (predominant research fields: gastrointestinal infections n=11, healthcare-associated and antibiotic-resistant infections n=10, hepatitis n=10). 48 % (40/84) of participants had experience as medical doctor. The three top ranked data categories were microbiology and parasitology, experimental data, and medication (53%, 52%, and 47% of maximal points, respectively). The most relevant data items for these categories were bloodstream infections, availability of biomaterial, and medication (88%, 87%, and 94% of maximal points, respectively). The ranking of requests of data integration and exchange is diverse and depends on the chosen measure. However, there is need to promote discipline-related digitalization and data exchange.

**Keywords.** Data integration, minimum data requirement, infectious diseases, survey

## 1. Introduction

The Medical Informatics Initiative (MII) of the German Federal Ministry of Education and Research (BMBF) is progressing rapidly in its effort to create data integration centers (DIC) allowing collection and sharing of real-life patient data in German university hospitals [1]. Within the MII, a core data set is developed to increase interoperability between datasets of different consortia. It is developed by experts in the MII interoperability working group for enabling data queries across consortia with a focus on relevance for medical and supply research, presence of interoperability standards and relevance for MII consortia use cases [2]. Two of the MII consortia support use cases related to infection research. However, these only cover very specific and small datasets that will not be of general use for infection disease (ID) researchers [3, 4].

The objectives of the German Infectious Disease Data Exchange (iDEx) project are to provide input of ID researchers in the process of data sharing in Germany and to identify minimum data requirements to support relevant use cases in ID research. Furthermore, these results could be used for future collaborations in the next round of the MII extension modules.

## 2. Methods

In consultation of the iDEx working party, we developed an online survey aiming to identify a ranking of requests for data integration and exchange. Individual needs and requests from different ID research fields were included into the survey. The survey was designed in three sub-surveys. The structure of the survey is displayed in **Figure 1**. In the following, the iDEx survey is described in detail.

### 2.1. Online survey

The online survey was designed by three separate sub-surveys including (i) a ranking of data categories, (ii) ranking of relevance of specific data items, and (iii) determination of available data collection systems and interoperability standards.

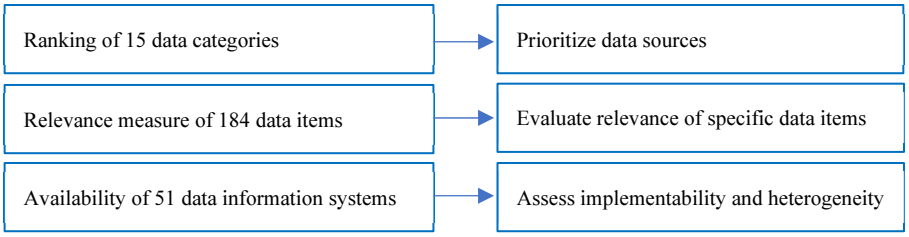


Figure **1.** Structure of the iDEx survey and output of sub-surveys.

The ranking for data categories was implemented as list, where participants were instructed to sort in a descending order the relevance of data categories. When data categories were not needed for one's particular research, the participant could leave categories unsorted. Participants were then asked to rank specific data items within the sorted categories. To compromise between the required level of detail and the expected longanimity of the participants, we chose a moderate richness of detail, e.g. HIV would be a data item in the virologic category instead of HIV-1-RNA, HIV-2-RNA, and HIV-1 Sequence, etc. The relevance of data items was determined using the Likert scale (strongly agree, agree, neutral, disagree, and strongly disagree).

For each data category, the participants were also asked, which specific data items are so important for them that no sensible research would be possible without that information, expecting e.g. HIV researchers to select CD4 and HIV-RNA, two basic required descriptors of disease stage and treatment success. This question was added a) to check completeness of extracted and collected information for the interview and b) to further assist in the prioritization of the most possible comprehensive core dataset.

Participants not affiliated with the German Center for Infectious Diseases (DZIF) were excluded, since the survey was intended to advise a DZIF IT-agenda for the most important next steps. Participants who did not complete ranking of categories or participated twice were excluded. All computations were based on relative data and excluded missing (not filled cells) information.

In the first sub-survey, the 15 data categories were ranked based on a rank-sum of $16 - x$ and normalized to the maximum achievable score based on number of participants. Unsorted categories were counted as 0. Additionally, the median and interquartile range of rank-sum for each category, the frequency of highest ranked category, and the frequency of votes where each category where irrelevant/unsorted.

In the second sub-survey the sum of Likert scale, defined as 5 to 1, were computed and normalized for each data item. The top most and lowest ranked data item for each category and ranking of data items for each category was computed. The distribution of data items which were absolutely necessary was computed.

A link to the iDEx survey was distributed via a DZIF mailing list with 795 email addresses (all DZIF affiliated researchers and project managers and secretariats). A reminder to participate in the survey was sent. All participants were asked to specify their

thematic translational unit (TTU) and thematic infrastructures (TI) affiliation, and experience.

### *2.2. Methods for data analysis*

The survey was developed and conducted via an established online platform (ClinicalSurveys, Questback GmbH, Cologne, Germany). Analysis was performed using Excel (Microsoft Corporation, Redmond, Washington, USA, version Office 2019, 2019) and R (R Development Core Team, Vienna, Austria, version 3.5.2, 2019) with package *dlyr*.

## 3. Results

In total, 84 participants have completed the online survey. Thereof, eight participants completed only the first sub-survey (ranking of categories). 48% (40/84) of participants had experiences as medical doctor, 56% (47/84) as data scientist, 70% (59/84) as lab scientist, and 62% (52/84) as project manager. The most commonly represented ID research fields were gastrointestinal infections with 13% (11/84), healthcare-associated and antibiotic-resistant bacterial infections and hepatitis both with 12% (10/84), tuberculosis with 10 % (8/84), and human immunodeficiency virus (HIV) with 8% (7/84). The three most represented sites were Munich (25/84; 30%), Cologne (13/84; 15%), and Hamburg (9/84; 11%). The results of the first sub-survey (ranking of data categories) are shown in **Table 1**. The ranking of each data item for each category is displayed in **Table 2**.

We identified that bloodstream infection (15/84; 18%), patient survival (13/84; 15%), dosage of medication (12/84 14%), temperature (11/84; 13%), agent (10/84; 12%), microbiome (10/84; 12%), antibiotics (9/84; 11%), comorbidities (9/84 11%), country of origin (8/84; 10%), biomaterial availability (8/84; 10%), genomic data single bacteria (8/84; 10%), and medication start/end date (8/84; 10%) were for more than 10% of participants absolutely necessary for research. **Table 3** gives an overview of the data information systems at sites of participants, which were present at minimum three sites.

## 4. Lessons learned (Discussion)

There is almost universal high demand for a digital agenda of ID researchers, irrespective of the field of research, rank of scientist, and translational focus. The median data requests focused on pathogen data, experimental and OMICS data, as well as clinical data. High diversity of requested data items was present and expressed in different rankings based on different measures of relevance. The top most data categories based on the frequency of highest ranked category were movement data (e.g. interaction data), diagnoses, and medication. In comparison, these categories were ranked 13, five, and three, respectively in the absolute ranking score. We identified contradictory results. For instance, the data category which was most often ranked as highest was movement data. However, movement data had also the second highest frequency considered as irrelevant. Furthermore, the distribution of the ranking scores differed marginally. These examples show, that there is not one solution in ranking of the request of data integration and exchange. It is not easily possible to delimit certain preferred categories.

**Table 1.** Ranking of data categories based on different importance measures. The ranking score is computed in defining the highest ranked category as 15 points and consecutive categories as 15 – 1.

| Category | Ranking score, max = 1,260 | Ranking score, median (IQR) | Frequency highest ranked, max = 84 | Frequency irrelevant, max = 84 |
|---|---|---|---|---|
| Microbiology and parasitology | 665 (53 %) | 9 (13 – 3) | 7 (8 %) | 19 (23 %) |
| Experimental data[a], biomaterial | 658 (52 %) | 9 (13 – 1) | 7 (8 %) | 20 (24 %) |
| Medication | 596 (47 %) | 9 (13 – 0) | 9 (11 %) | 27 (32 %) |
| Laboratory[b] | 546 (43 %) | 7.5 (12 – 0) | 4 (5 %) | 29 (35 %) |
| Diagnoses[c] | 542 (43 %) | 7.5 (11 – 0) | 3 (4 %) | 25 (30 %) |
| Virology | 528 (42 %) | 6 (12 – 0) | 9 (11 %) | 34 (40 %) |
| Pathology | 446 (35 %) | 0 (12 – 0) | 6 (7 %) | 46 (55 %) |
| Sociodemographic data, anamnesis[d] | 424 (34 %) | 0 (11 – 0) | 7 (8 %) | 47 (56 %) |
| Health economics | 418 (33 %) | 0 (12 – 0) | 7 (8 %) | 51 (61 %) |
| Regular ward clinical course[e] | 416 (33 %) | 2.5 (11 – 0) | 3 (4 %) | 38 (45 %) |
| Outpatient clinical course[e] | 414 (32 %) | 2.5 (9 – 0) | 1 (1 %) | 41 (49 %) |
| Imaging methods | 403 (32 %) | 0 (11 – 0) | 2 (2 %) | 49 (58 %) |
| Movement data[f] | 342 (27 %) | 0 (9 – 0) | 13 (15 %) | 56 (67 %) |
| ICU clinical course[g] | 285 (23 %) | 0 (8 – 0) | 3 (4 %) | 53 (63 %) |
| Foreign material[h] | 226 (18 %) | 0 (0 – 0) | 3 (4 %) | 65 (77 %) |

ICU: intensive care unit, [a] genomic data, microscopy imaging, microbiome, transcriptome, etc.; [b] biochemistry, hematology, hemostaseology, drug levels, urine chemistry, etc.; [c] comorbidities, severity of disease/staging, data from Eurotransplant, ischemia periods, etc.; [d] place of residence, origin, hospital stays, recovery process, nursing, etc.; [e] vital parameters, respiratory parameters, outcome, clinical scores, etc.; [f] patient-patient-contact, Global Positioning System data, interaction data, travel, etc.; [g] vital parameters, mechanical ventilation, catecholamines, dialyses, respiratory parameters, outcome, medication, clinical scores, etc.; [h] foreign body, joint replacement, central venous catheter, heart valves, pacemaker, etc.

**Table 2.** Relevance measure of data items for each category. The relative points (rel. points) over the maximum points (always "strongly agree") are computed.

| Category | Best ranked data item | Rel. points | Worst ranked data item | Rel. points |
|---|---|---|---|---|
| Microbiology and parasitology | Bloodstream-infections | 88 % | Parasitological results | 73 % |
| Experimental data, biomaterial | Data on biomaterial availability | 87 % | Genomic data, sequences of tumors | 60 % |
| Medication | Agent | 94 % | Brand name | 71 % |
| Laboratory | Blood count | 89 % | Bone marrow cell counts | 58 % |
| Diagnoses | Stage/grade/severity of main diagnosis | 92 % | Transplantation details | 80 % |
| Virology | Hepatitis viruses and HIV | 86 % 86 % | Hemorrhagic fever viruses | 71 % |
| Pathology | Infection-related results | 90 % | Images | 73 % |
| Sociodemographic data, anamnesis | Symptoms | 91 % | Employment status | 68 % |
| Health economics | Billing-relevant data | 80 % | Health insurance data | 72 % |
| Regular ward clinical course | Patient survival | 95 % | HCT-CI | 61 % |
| Outpatient clinical course | Patient survival | 96 % | HCT-CI | 62 % |
| Imaging methods | CT/MRI findings | 83 % | Bone density | 48 % |
| Movement data | Travel anamnesis | 85 % | GPS data | 69 % |
| ICU clinical course | Patient survival | 95 % | ABSI | 60 % |
| Foreign material | Central venous catheters | 78 % | Pacemaker | 70 % |

**Table 3.** Overview of data information systems at sites of participants. Each site counted once.

| Data information system | Frequency, n = 14 |
|---|---|
| SWISSLAB (laboratory) | 8 |
| Orbis (HIS) | 7 |
| Lauris (laboratory) | 7 |
| CentraXX (biobanking) | 6 |
| PathoPro (pathology) | 4 |
| Meona (clinical data) | 3 |
| i.s.h.med (HIS) | 3 |
| HyBASE (hygiene) | 3 |

HIS: hospital information system

Considering the distribution of data items, which are absolutely necessary for research, it comes clear, that data items from different categories are necessary to solve research questions. Consequently, the integration of different data information systems would be necessary to extract medical data for research. A step-wise prioritized integration of data information systems could be used to extend and thereby improve local databases until a system-wide data integration is implemented.

Regarding the results of the ranking of relevance of different data items, it comes clear that the importance of data items in data categories (mostly present in the same data information system) differs. This grading could be used to prioritize the exchange of certain data items, when the respective data information system (mostly represented as data category) is integrated. This means, the results could be used to define an order for the introduction of interoperability standards (e.g. terminologies) for certain parameters. This prioritization is particularly helpful for the current introduction of SNOMED CT in Germany (initially for its use in the MI initiative) [5]. Due to the complexity of this international reference terminology, for example, a focus on relevant applications is needed for the generation of validated subsets. This iDEx study facilitates such a focus.

One could hypothesize, that the ranking of data categories reflects the distribution of researchers in different ID research fields, as certain research questions require the use of certain data information systems. Therefore, the results of this study should help identifying overlapping needs of different ID research fields. We included in the online survey many data items. The complexity of the online survey could create a bias in data items, which were listed at the end of the survey. Additionally, the information of available data information system could be biased, as this information could not always be available for participants. Small research fields (TTUs/TIs) could be underrepresented. However, after discussion within the iDEx working party this is seen as a neglectable factor.

To our knowledge, this is the first study identifying interests for digitalization of a specific research topic in a broad national scale. The requests of researchers with different experience were included in contrast to expert opinions which are often used for consensus-building in committees [2].

To our knowledge, the identified preferred data categories (pathogen, experimental data, and detailed clinical data) will not be included (so far) in the core data set developed by the MII. Reasons are described as less relevance for research or MII use cases, availability and implementability, or interoperability compared to other data categories/information systems [2]. We are confident, that especially pathogen and detailed clinical data could be from high relevance for different DZGs, as e.g. the German Centre for Cardiovascular Research or German Centre for Diabetes. However, in other

disciplines, we would expect heterogeneous results as in our study, as different perspectives of clinical, translational, and experimental research are mostly also present. In addition, the common data integration of different data sources can be improved in future by the need for a network for COVID-19 data.

## 5. Conclusion

The ranking of requests of data integration and exchange is diverse and depends on the chosen measure. There is a need to promote discipline-related digitalization and data exchange. Only with available data sets, which are comprehensive enough, research could be raised to a next level adapted to the era of digitalization.

## Authors' contribution

J. Vehreschild (JV), C. Jakob (CJ), M. Stecher (MS), S. Fuhrmann (SF), G. Anton, M. Behnke, U. Behrends, M. Boeker, S. Castell, H. Demski, M. Diefenbach, J. Falgenhauer, M. Fritzenwanker, P. Gastmeier, M. Gerhard, S. Glöckner, M. Golubovic, B. Gunsenheimer Bartmeyer, J. Ingenerf, R. Kaiser, M. Körner, W. Loag, A. McHardy, E. Molitor, U. Nübel, M. Pritsch, M. Ramharter, S. Rieg, J. Rupp, D. Schindler, D. Schwudke, C. Spinner, B. Stottmeier, M. Vehreschild, M. Willmann designed the study. JV initiated the project. MS, SF, and CJ conducted the interviews. SF, S. Heinen, and S. Wingen-Heimann extracted the data. CJ and MS analyzed the data. CJ, MS, SF and JV interpreted the data. CJ, MS and JV drafted the manuscript. The iDEx group revised the manuscript critically for important intellectual content.

## Acknowledgements

We thank the German Center for Infection Research for financial support. We thank all participants of the survey for the contribution in this study.

## Conflict of Interest

The authors state that they have no conflict of interest.

## References

[1] Semler, S.C., F. Wissing, and R. Heyder, *German medical informatics initiative.* Methods of information in medicine, 2018. **57**(S 01): p. e50-e56.

[2] Ganslandt, T., et al. *Der Kerndatensatz der Medizininformatik-Initiative: Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene*. in *Forum der Medizin-Dokumentation und Medizin-Informatik*. 2018.

[3] Haarbrandt, B., et al., *HiGHmed–an open platform approach to enhance care and research across institutional boundaries.* Methods of information in medicine, 2018. **57**(S 01): p. e66-e81.

[4] Winter, A., et al., *Smart medical information technology for healthcare (SMITH).* Methods of information in medicine, 2018. **57**(S 01): p. e92-e105.

[5] Bundesministerium für Bildung und Forschung (2020-03-20): Digitalisierung: Medizinische Daten sprechen zukünftig eine gemeinsame Sprache. URL: https://www.bmbf.de/de/digitalisierung-medizinische-daten-sprechen-zukuenftig-eine-gemeinsame-sprache-11140.html [20-04-04]