SCALING UP THE BIG HEALTH DATA ECOSYSTEM: ENGAGING ALL STAKEHOLDERS!

Dipak Kalra PhD, FRCGP, FBCS, FIAHSI

The European Institute for Innovation through Health Data

Abstract

There is now an urgent need to scale up our collective capability to learn insights from health data, to improve patient care pathways and health services, to ensure that public health measures and strategies are underpinned by real time evidence, and to accelerate research such as the development of drugs, vaccines and AI algorithms. Europe is investing within and across countries in research infrastructures to enable this scaling up, most frequently through federated architectures. The latest development is the plan from the European Commission to create a European Health Data Space. However, any architecture to combine data or to run distributed queries is critically dependent upon the data being held or mapped to a standardised form (structurally and semantically). Standards exist to achieve this, although more stakeholder engagement is needed in defining practical clinical models and value sets, but the real adoption of interoperability is disappointing and needs further incentivisation and investment. Data quality is another concern that can only be improved if there is awareness that this is important, a willingness to invest and a recognition that many stakeholders need to become motivated to improve quality. Scaling up the uses of data also means involving new actors such as industry. Societal trust is a vital prerequisite for enabling novel uses of data. Transparency is a critical success factor for trust. Data access governance rules must be developed through open public consultation. The bodies who make access decisions must publish information about the data accesses they have permitted. For the public to be on board they have to understand much more than most people do about the nature of health data, how it can be used for the benefit of society and what safeguards protect them when the data are used.

Keywords: electronic health records; clinical research; data architectures; information governance; learning health systems

Kalra D. JISfTeH 2020;8:e16(1-5).
DOI: https://doi.org/10.29086/JISfTeH.8.e16

Introduction

There is now a compelling global need for all health and research stakeholders to collaborate in accelerating our capability to learn from health data at scale, and to translate that learning into diagnostic and treatment innovations, care pathway transformation and novel digital solutions. The COVID-19 pandemic has shown how hard it is for us to collect new data sets to a high quality, to be able to share data across borders (and even within borders) and to be able to use it for strategic insights to enable more accurate and better targeted public health and health system responses. This has been possible in example areas, such as a global co-operative in multiple sclerosis, the Multiple Sclerosis Data Alliance,[1,2] and studies starting to be published by the Observational Health Data Sciences and Informatics (OHDSI) community.[3] It is vital that we use the lessons highlighted by COVID-19 to accelerate those critical success factors to enable us to better respond to any future unexpected scenario, as well as to improve how we handle our current health and care crisis: long-term conditions and multimorbidity.

Health systems are challenged by increasing multimorbidity, due to our ageing population,[4] and struggle with delivering their complex care management needs.[5] More than half of all older people have at least three chronic conditions, and a significant proportion has five or more.[6] Poorly managed multimorbidity may increase the risk of disease complications and vulnerability due to acute deteriorations, for example hospitalizations, falls and death.[7] Higher healthcare resource consumption in these patients is not only because of the accumulation of chronic health conditions but also because of interactions and synergies among health conditions present within an individual.[8] Our knowledge about these interactions is limited. For example, the C3-Cloud European project needed to rely heavily upon clinical judgement to work out how best to optimise a multi-condition care pathway when the starting point was four single disease clinical guidelines that had been developed in isolation.[9] However, there will probably be tens of thousands of patients with some combination of four common diseases from whom we could learn which treatments and other care pathway elements had been the most effective and safe. Why

are we not learning this from our data already?

One of the important challenges with closing our knowledge gaps is the need for large-scale data, so that we have sufficient patient numbers to examine different multimorbidity patterns, to stratify patients into biomarker-specific profiles that may respond best to different interventions and to further develop our understanding and treatment of rare diseases. Large scale data is sometimes the only way to detect small effect sizes, as recently demonstrated for first line hypertension therapy by the OHDSI community as part of the Longitudinal Examination to Gather Evidence of Neurodegenerative Disease (LEGEND) study.[10]

We now have important initiatives that are scaling up our ability to connect and analyse multiple data sources. These are increasingly favouring a federated rather than a centralised architecture. There are several advantages of a federated model: the data sources each remain their "source of truth" which means there is a single place where updates and version management are handled; each data source retains autonomy over the purposes and parties for data reuse that they will endorse; there are a fewer issues about data ownership and cross-jurisdictional data transfers. There are novel techniques that not only encrypt distributed queries and the result sets but permit federated queries to be performed on data sets that remain encrypted throughout the analysis.[11] Personal data can therefore remain strongly safeguarded even at the nodes that are performing the queries throughout the federation. Public concern might therefore be lower, although this topic needs more careful investigation.

Probably the largest European projects to tackle the design, implementation and scale up of federated research networks have been the Innovative Medicines Initiative projects, the European Medical Informatics Framework project (EMIF) and the European Health Data and Evidence Network (EHDEN).

The EMIF project undertook five and a half years of R&D to design and implement a platform and tools to conduct research across a distributed network of European health data sources. EMIF's aim was to establish the mechanisms to accelerate the scaling up of big data research, by designing and implementing a multi-component architecture to capture and cascade research queries to multiple connected data sources.[12] Each data source was invited to create a shadow data warehouse containing only the data that the source was willing to make available through the EMIF federation, mapping it to the Observational Medical Outcomes Partnership (OMOP) common data model.[13] The EMIF results also included establishing a data catalogue to enable data sources to be discovered and characterised, so that a researcher could determine its suitability for their research study, and a code of practice that data sources and research users must adhere to in order to ensure mutual respect and recognition, and to protect data privacy. A successor project, EHDEN, is now scaling up the EMIF results, underpinned by the OHDSI architecture.[14]

Real-world data, especially from hospitals, is also proving valuable to help optimise the design and conduct of clinical trials. The re-use of electronic health records can increase and speed up patient recruitment into clinical trials, making trials more likely to complete successfully and on time.[15] The Electronic Health Records for Clinical Research (EHR4CR) project developed the first EHR-vendor neutral platform to federate multiple hospital EHRs in order to enable trial protocol design to be based more accurately on real patient numbers rather than estimates, and then to facilitate the recruitment of eligible patients by hospitals participating in a trial.[16] The platform design has now been successfully commercialised.[17] A successor project, Electronic Health Records to Electronic Data Capture (EHR2EDC) has implemented and validated a pipeline to enable the EHR data on a trial participant (after consent) to be transferred into the clinical trial EDC system to avoid duplicate data entry efforts and errors.[18]

There is now great interest across many organisations in the plans announced by the European Commission for a series of common European data spaces.[19] This overall strategy is illustrated in Figure 1.

The data input sources, the potential users and the governance environment for the European Health Data Space (EHDS) are still in development. There are several existing European data networks, including the eHealth Digital Service Infrastructure (eHDSI) that shares patient summaries and electronic prescriptions across Europe, the European Reference Networks (ERNs), the networks established between regulatory agencies across Europe known as DARWIN (Data Analysis and Real World Interrogation Network) and the life sciences research infrastructures such as ELIXIR and BBMRI (Biobanking and Biomolecular Resources Research Infrastructure), all of which might have connection points to the EHDS. National health and research networks, such as those in Germany, France, Scandinavia, are also candidates for connection. Several key stakeholder groups, especially industry, might be data providers to this space, as well as being possible data users alongside public health agencies. It is unclear at present whether the EHDS will be mainly federated, with little centrally health data, or will be primarily a centralised data store of high-value data sets extracted from these networked infrastructures.

However, whether a federated or centralised architecture is used by the EHDS and by other data resources, our ability to scale up the analysis of health data will stumble unless the data are held in standardised forms. We have standards for the technical communication of data "down the wire" and there are common data models like OMOP for mapping data into a federation-ready form. However, our routinely collected clinical data, mostly in EHR systems, still supports standards to a limited extent. Although we have high-level information model standards and terminology standards (do we have too many?), the problem is putting these together

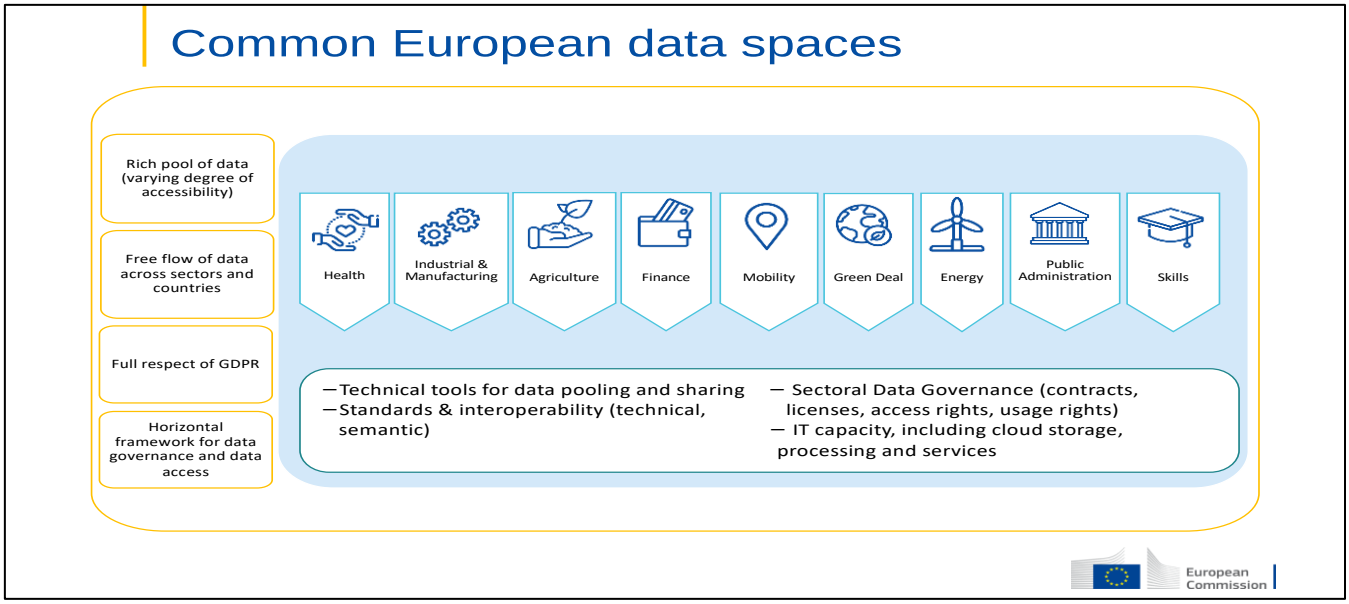

Figure 1. Common European data spaces. Source: the European Commission.

into practically usable and digestible clinical models and value sets to encourage diverse specialisms and professions within healthcare to collect and share their data in the same form. If we have such clinical data standards, we can link them to decision support and analysis queries in order to get more reliable results. The need for this area of "practical standardisation" has been stated for many years,[20] but we still lack adequate investment in building communities of practice who can specify these finite "building blocks" through consensus, and build the momentum for their widespread adoption for data capture as well as interoperability.

One example of a more focused and practical ambition has been to standardise and promote the adoption of an international patient summary. Building on the two parallel initiatives towards a standardised health summary for patients, the EU sponsored Trillium Bridge project (2013-2015) compared patient summary standards and specifications in Europe and the United States and demonstrated the technical feasibility of exchanging electronic health record summaries across the Atlantic in the context of emergency or unplanned care abroad. Its successor project, Trillium II (2017-2019) extended the use cases for an international patient summary and demonstrated its potential value.[21] Trillium II championed international standardisation, and this is now embedded within HL7 and ISO work plans to publish an International Patient Summary standard.[22,23]

More work is needed to define other high priority data sets behind which multi-stakeholder efforts can be focused, for example that developed by the EHR2EDC project: a dataset that offers the best real-world data utility for clinical trials (to be published in late 2020).

Even if we have architectural solutions and widespread standards adoption, we will still fail to generate trustworthy inferences from data unless the quality of that data is good enough. As an example of this problem, Doods et al demonstrated that even basic measurements like body weight can be missing from the EHRs of patients with important health conditions where this would be expected.[24] If missing data can lead to serious healthcare consequences such as medication dosing errors,[25] then one would expect the quality of EHR data to also risk incorrect research analysis results. This has prompted organisations like the European Institute for Innovation through Health Data (i~HD) to establish a data quality assessment and improvement programme, to help hospitals to raise the quality of their data in order to participate more successfully in research as well as to improve their ability to learn from their own data to improve care.[26] Data quality not only means minimising incomplete documentation but ensuring that the data values that are entered are consistent with the data items being filled, comply with any implemented data dictionary, and that the values are sensible in the context of the patient and of that patient population.

The assurance of societal trust is also a vital prerequisite to scaling up the range of actors and purposes for which health data may be used. There are plenty of examples over the past 20 years where attempts to ski club data use, data sharing and data networks have failed because of a public backlash. The challenge we face is that the further the purposes and actors are from a patient's place of familiarity (the health services and the healthcare professionals they know), the harder it is for people to be comfortable about the uses being made of their data, the parties who were making that use, how their identity and interests are being safeguarded, and whether they support those uses of the data (see Figure 2).

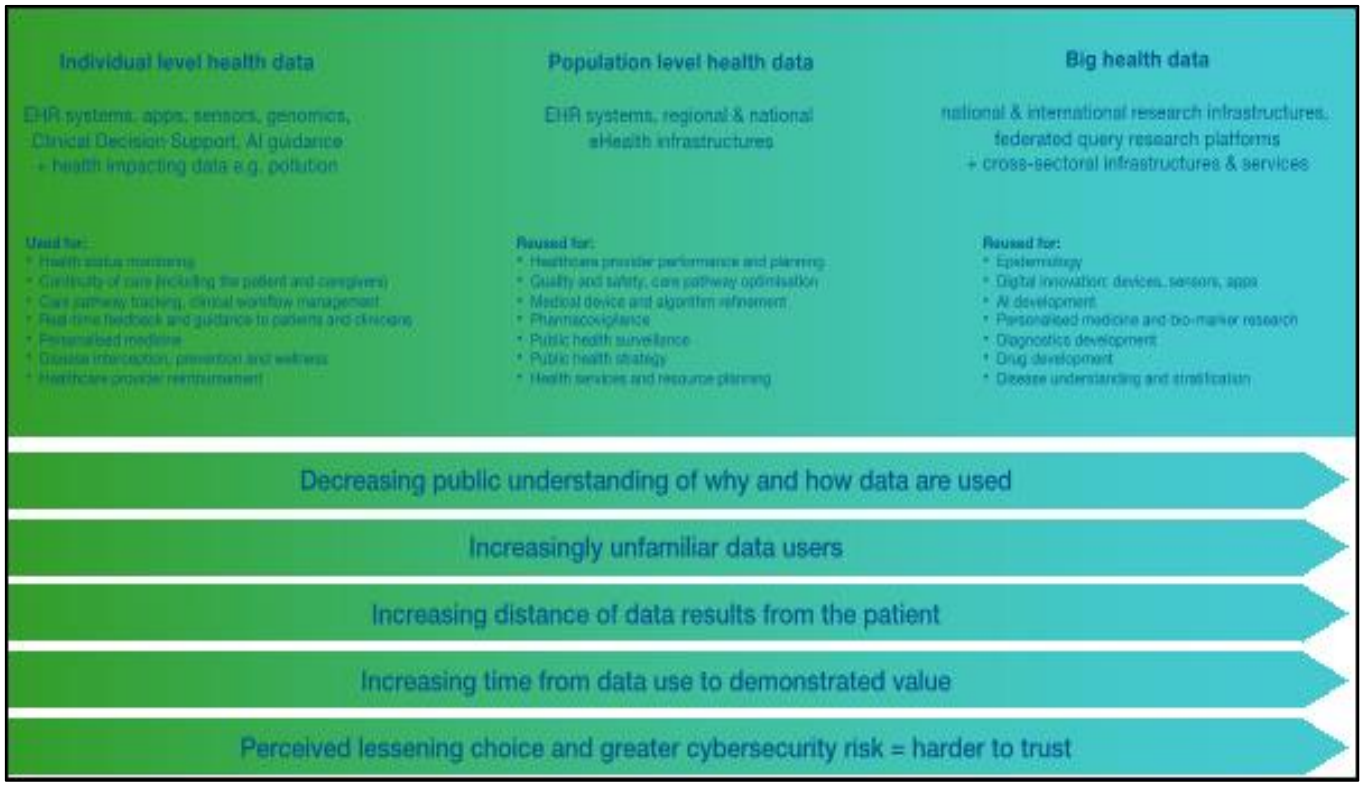

Figure 2. The challenge with gaining public acceptance of health data reuse.

A substantial public education programme is needed to help people to understand why it is important that health data be widely used, the benefits of this use and the safeguards that can be adopted. The Data Saves Lives initiative is spearheading a new public awareness campaign on this across Europe.[27] To complement this, organisations who make use of health data need to be bound by practices and codes that ensure public trust is well placed. The governance framework for the EHDS is hoped to include a European code of conduct for health data use, which has the prospect of increasing public trust in how their health data are used.

When we think about the use of health data it is vital not to forget that patients and healthy citizens are not only data creators, contributing to the learning that can be made by others. They must themselves be empowered to make use of their own data through apps, sensors and smart feedback loops. We will increasingly see people getting this real-time feedback, sometimes comparing their data with others in a similar community, being offered localised and personalised alerts or help with setting goals and reaching targets.

The more that we bring patients and healthy citizens inside the learning loop with data, the more they will understand about the power of data and the importance of powering up the learning health system.

Corresponding Author:
Dipak Kalra
The European Institute for Innovation through Health Data
E-mail: dipak.kalra@i-hd.eu

Conflict of interests: the author declares no conflicts of interest.

References

1. The Multiple Sclerosis Data Alliance (MSDA). (2020). Available at: https://msdataalliance.com accessed on 6 September 2020.
2. Peeters LM, Parciak T, Walton C, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. *Mult Scler J* 2020;26(10):1157-1162. DOI:10.1177/1352458520941485
3. Observational Health Data Sciences and Informatics. (2020). COVID-19 Updates. Available at: https://www.ohdsi.org/covid-19-updates/ accessed on 6 September 2020.
4. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;380:37–43. DOI: 10.1016/S0140-6736(12)60240-2

5. Vetrano DL, Calderón-Larrañaga A, Marengoni A, et al. An international perspective on chronic multimorbidity: approaching the elephant in the room. *J Gerontol A-Biol.* 2018;73(10):1350–1356. DOI: 10.1093/gerona/glx178
6. Luppi F, Franco F, Fabbri BBL. Treatment of chronic obstructive pulmonary disease and its comorbidities. *Proc Am Thorac Soc* 2008;5(8):848-856. DOI: 10.1513/pats.200809-101TH
7. Calderón-Larrañaga A, Vetrano DL, Ferrucci L et al. Multimorbidity and functional impairment–bidirectional interplay, synergistic effects and common pathways. *J Intern Med* 2019;285:255–271. DOI: 10.1111/joim.12843
8. Mercer S, Salisbury C, Fortin M. ABC of Multimorbidity. New York: Wiley, 2014.
9. The C3-Cloud project. (2020). Available at: https://c3-cloud.eu/solutions-health-sector accessed on 6 September 2020.
10. Schuemie MJ, Ryan PB, Pratt N et al. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Assoc* 2020;27(8):1331-1337. DOI: 10.1093/jamia/ocaa103
11. Paddock S, Abedtash H, Zummo J, Thomas S. Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine. *BMC Med Inform Decis Mak* 2019;19(1):255. DOI: 10.1186/s12911-019-0983-9
12. Lovestone S. EMIF Consortium. The European medical information framework: A novel ecosystem for sharing healthcare data across Europe. *Learn Health Sys* 2019;4(2):e10214. DOI: 10.1002/lrh2.10214
13. Observational Health Data Sciences and Informatics. (2020). (OMOP) Common Data Model. Available at: https://www.ohdsi.org/data-standardization/the-common-data-model/ accessed on 6 September 2020.
14. The European Health Data & Evidence Network. (2020). Available at: https://www.ehden.eu accessed on 6 September 2020.
15. Dugas M, Lange M, Muller-Tidow C, Kirchhof P, Prokosch HU. Routine data from hospital information systems can support patient recruitment for clinical studies. *Clin Trials* 2010;7(2):183–189. DOI: 10.1177/1740774510363013
16. De Moor G, Sundgren M, Kalra D et al. Using electronic health records for clinical research: The case of the EHR4CR project. *J Biomed Inform* 2015;53:162-173. DOI: 10.1016/j.jbi.2014.10.006
17. Insite. (2020). InSite the Largest European Live Clinical Data Network. Available at https://www.insiteplatform.com accessed on 6 September 2020.
18. i~HD. (2020). Enriching knowledge and enhancing care through health data. Available at: https://www.i-hd.eu/index.cfm/r-d-and-collaborative-projects/research-projects/ehr2edc/ accessed 14 December 2020.
19. European Commission. COM(2020) 66: A European strategy for data. (2020). available at: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf accessed on 6 September 2020.
20. Stroetmann VN, Kalra D, Lewalle P, et al. Semantic Interoperability for Better Health and Safer Healthcare: Research and Development Roadmap for Europe: Semantic HEALTH Report. (2009). DOI: 10.2759/38514 Available at: https://op.europa.eu/en/publication-detail/-/publication/9bb4f083-ac9d-47f8-ab4a-76a1f095ef15 accessed on 6 September 2020.
21. Chronaki C, Estelrich A, Cangioli G, et al. Interoperability standards enabling cross-border patient summary exchange. *Stud Health Technol Inform* 2014;205:256-260. PMID: 25160185.
22. Hl7FHIR. (2020). International Patient Summary Implementation Guide. Available at: http://www.hl7.org/fhir/uv/ips/2018May/ accessed on 6 September 2020.
23. ISO. (2020). ISO/FDIS 27269 Health informatics — International patient summary. Available at: https://www.iso.org/standard/79491.html accessed on 6 September 2020.
24. Doods J, Botteri F, Dugas M, Fritz F. A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials* 2014;15(1):18. DOI: 10.1186/1745-6215-15-18.
25. Hirata KM, Kang AH, Ramirez GV, Kimata C, Yamamoto LG. Pediatric weight errors and resultant medication dosing errors in the emergency department. *Pediatr Emerg Care* 2019;35(9):637–642. DOI: 10.1097/PEC.0000000000001277
26. i~HD. (2020). Data Quality Champion Programme. Available at https://www.i-hd.eu/index.cfm/services/health-data-quality/data-quality-champion-programme/ accessed on 6 September 2020.
27. Data Saves Lives. Available at: https://datasaveslives.eu/home accessed on 6 September 2020.