Infrastructure for the Learning Healthcare System: Centralized or Distributed?

Andrius Budrionis[1] and Johan Gustav Bellika[1,2]

[1] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
Andrius.Budrionis@ehealthresearch.no

[2] Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway

Abstract

The learning healthcare system ideas brought novel approaches to knowledge generation and adoption in healthcare. High focus on data created in routine practice highlighted the underexplored potential for quality improvement, cost reduction and personalized care. While the high-level goals promise major improvement in healthcare delivery worldwide, it is less clear how such practices should be implemented and maintained. Technical infrastructure to ensure data access and processing capabilities is the initial building block enabling agile learning from practice. Selection between distributed data processing network and centralized repository as a fundament for the learning healthcare system has implications to further development of the system and its utility. This paper contrasts two architectural patterns and highlights their advantages and limitation to serve data access and processing needs in the learning healthcare. It provides comparative insights in decision making and assists in selecting an approach for implementation.

Keywords

Learning healthcare systems, centralized, distributed, architecture, data reuse, infrastructure, primary care.

1 INTRODUCTION

Evidence based medicine (EBM) has left its footprint in modern healthcare by developing new research methodologies, building novel knowledge bases to be implemented in clinical care and increasing scientific publication standards. However, overwhelming amount of evidence in a form of scientific papers and slow adoption of knowledge in practice were identified as major limitations of EBM (Greenhalgh et al., 2014). It is shocking that the average duration of implementing research knowledge in clinical practice may span to 17 years (Morris et al., 2011). This finding gave an impetus for a novel approach to knowledge creation and adoption in healthcare – the Learning Healthcare System (LHS) (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007).

Since the first time mentioned in 2007, LHS attracted major attention in academia worldwide. The conceptual ideas were further developed with increasing expectations that the continuous loop of data-knowledge-clinical practice will change knowledge generation in medicine as we know it. Faster progression of research knowledge into practice, improved adaptation to individual patient needs, delivering personalized care, better balance in resource and costs utilization are major promises making LHS a preferred future direction for healthcare (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007). However, after ten years of global effort, we are not much further compared to where we started. Even though the limitations in healthcare delivery are often clear, implementing changes still takes long. A recent review on LHS implementations confirmed it by identifying a relatively low number of attempts to implement LHS in practice. While the expectations in LHS are high, little is done to actually adopt LHS practices in reality (Budrionis and Bellika, 2016).

Lacking guidance on the infrastructure development was identified in the review. It is not yet clear whether LHS infrastructure should work as a distributed computation network (Figure 1A), analyzing health data close to real-time or a centralized repository (Figure 1B) accumulating data from various sources and making it ready for statistical analyses. The reviewed LHS instances divided almost equally when selecting their approach to the technical architecture (6 – distributed, 7 – centralized) (Budrionis and Bellika, 2016). Since both approaches have their strengths and weaknesses, this paper aims to compare them providing comprehensive argumentation for future development to enable an informed decision when choosing between the centralized and distributed LHS architectures. This paper looks into fundamental factors differentiating both approaches and present a comparative analysis based on predefined metrics.

Paper from the 15th Scandinavian Conference on Health Informatics SHI2017, Kristiansand, Norway, 29 - 30 August.
Conference Proceedings published by Linköping University Electronic Press.

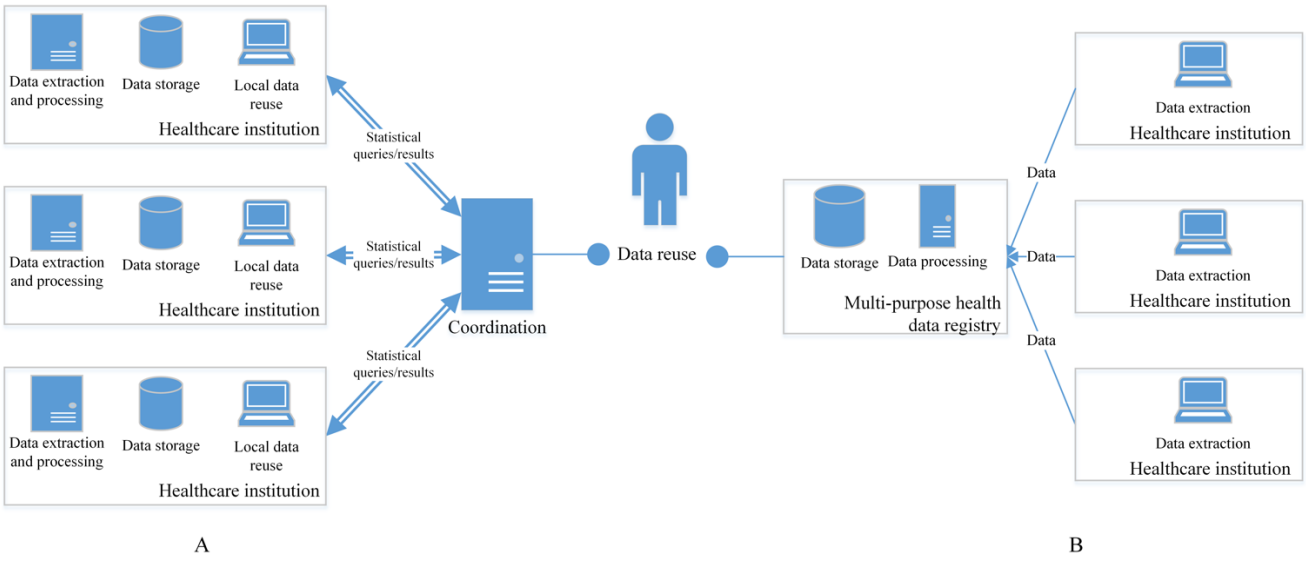

Figure 1 infrastructure for the LHS. A - distributed, B - centralized

2 MATERIALS AND METHODS

To compare the distributed and centralized LHS infrastructures an analysis framework was established. It originates from personal experience in building a distributed LHS infrastructure in Norway and combines characteristics, which were identified as important in this process. Centralized infrastructure refers to clinical data repository or registry, accumulating large amounts of data from various sources (Figure 1B). For the purpose of comparability, we assume that centralized system has the same data refresh rate as distributed one and does not store patient identifiable information. Moreover, unlike the national health registries[1], centralized LHS infrastructure is assumed to accumulate broad spectra primary care data without focusing on a certain use case (for instance, cause of death registry).

Distributed architecture functions as a network of computational nodes processing patient data locally in healthcare institutions and only sharing final results (Figure 1A).

Seven comparison areas were analyzed, giving a deeper insight in both architectural patterns. Data access and timeliness (1) was regarded as the fundamental piece of functionality fulfilling the main requirements for the LHS infrastructure. Capacity of statistical analyses (2) defines the potential of data reuse and ability to provide proof for behavior change in healthcare, while having sufficient measures to ensure patient privacy and data security (4). Having control over data (3) is of high importance, especially in primary care, where clinicians have an ownership relation to patient information they are collecting. Quality of data and its utility (5) in knowledge generation needs to be maximized together with the availability (6) to ensure sufficient feedback to the parties contributing to the overall system. Ability to integrate knowledge bases across borders (7) needs to be evaluated with regards to the legal frameworks and architectures for supporting the evolvement of the LHS. The aforementioned criteria define the comparison framework, which is explored in the remainder of this paper.

3 RESULTS

3.1 Data access and timeliness

Access to data is the main prerequisite enabling continuous LHS loop. Routine data is the source of knowledge for quality improvement and performance monitoring, therefore accessing it in a timely manner is a natural starting point when adopting LHS in practice. A traditional data access approach leveraging the long lasting experience is a centralized data storage. Clinical data is transferred and stored in a centralized repository at predefined time intervals, often as seldom as once every few years. Referring to the assumption made in Method section, data refresh rate is expected to be the

[1] https://www.fhi.no/en/hn/health-registries/cause-of-death-registry/about-the-national-health-registries/

same on both centralized and distributed architectures, therefore it is not highlighted as disadvantage.

On contrary to centralized data storage, distributed access to data implies no movement of data from the location it is collected. Data processing capacity is moved instead, implying access to close to real-time data. In other words, the infrastructure is a network of computational nodes, performing statistical computations on local data and aggregating the results. In this hypothetical case of equivalent data refresh rate, data access properties are comparable on both approaches. However, requirements for availability and reliability of the overall system are higher for distributed architecture to ensure sufficient coverage of data sources (discussed in section Data availability).

3.2 Capacity of statistical analyses

LHS is built on the idea of extracting knowledge from data. Therefore, the capacity of available statistical analyses is of high importance for achieving the goals of the LHS. Centralized data storage, in this case, does not have limitations with regards to methods, techniques and algorithms. Established data processing and analysis tools could be chosen based on personal preferences and results of the calculations can be reproduced at any time.

Distributed data analysis counts on distributed statistical processing algorithms, which are often referred to as research initiatives rather than established data processors (Asfaw Hailemichæl et al., 2015; Bellika et al., 2015; Yigzaw et al., 2013). The availability of supported statistical functions is currently limited, however, it is improving and more off-the-shelf tools will likely become available in the future. Major limitation of distributed data processing is the lacking ability to explore data and verify the correctness of computations. Moreover, reproducing results from previous computations may be complicated due to changes in data and availability of the processing nodes.

3.3 Control over data

Care providers have the responsibility of ensuring secure management of patient data. Norwegian primary care consists of a large number of relatively small GP offices, running their EHR systems locally or hosted by municipality. Therefore, GPs often have a special ownership relationship to the data they are collecting. Having all data locked up in a server room gives a feeling of control over sensitive information and helps implementing security measures. Regardless of the aforementioned attitude, patients are data owners, currently having limited or no control over their personal information accumulated by healthcare services.

Getting access to such data for research is often a difficult process highly dependent on the willingness of GPs to collaborate. Moving data from EHR to a research repository (registry) for a GP means losing control over patient information, which could later be used for various purposes (for instance, monitoring the performance of a certain GP). On the other hand, distributed data processing infrastructure ensures that patient data does not leave the institution it was generated and is only used in computations. Only the final product is then shared across institutions, protecting the privacy of both patient and GP. In this case GPs maintain the ownership of the data while enabling research and quality improvement activities. Moreover, distributed infrastructure ensures the ability to opt out of the network at the same time revoking access to data without involving infrastructure administrator or leaving any data trace.

3.4 Patient privacy and data security

Health data breaches are becoming a major threat to patient data security. Number of incidents and severity is increasing as reported in the American Health Information Breach Portal (U.S. Department of Health and Human Services, Office for Civil Rights, n.d.). Sixteen percent of reported incidents are caused by hacking or related to security of IT systems, however, they are responsible for health data losses of 79% of the population affected by incidents of handling health data from 2010 to 2016 (U.S. Department of Health and Human Services, Office for Civil Rights, n.d.). Similar trends are likely in other developed countries, highlighting that data security in healthcare sector is not managed well enough. This fact needs to be kept in mind when selecting the design approach for the LHS.

The level of sensitivity of health data needs to be assessed when deciding what data should be available for the LHS. Direct patient identifiers (personal number, phone number, etc.) can be replaced or hashed, however, reidentification risk still remains, especially in cases of rare diagnosis/symptoms and small patient communities.

Selection of data access and processing infrastructure plays a role in ensuring the security of patient data. If we define the risk as a product of likelihood and consequences, it is clear that centralized data storage, accumulating all data has higher security risk profile than distributed infrastructure. The consequences of a data breach in an infrastructure containing data for the entire population are catastrophic. An equivalent data breach in a distributed infrastructure implies getting access to every node in the system, containing a small fragment of the entire dataset stored centrally. The likelihood of compromising all nodes in the network is very small.

3.5 Data quality and utility

To preserve patient privacy, compromises on data quality and completeness need to be made. This process is more evident in a centralized architecture where access to the extracted data is available to the users. Data exploration step raises major security threats and may require information, which does not identify the patient directly, but could be used to find out the identity, to be excluded (for instance, zip code could be sufficient to identify a person having a rare condition). Relations between data variables could provide insights into the identity of the patient. Excluding all data, which potentially identifies the

patient (directly and indirectly) threatens the utility of the dataset.

The lacking direct access to raw data in a distributed architecture for data processing could be seen as an advantage from patient privacy and data quality point of view. Relatively few patient identifiers need to excluded, maximizing the utility and maintaining higher quality of the available data.

3.6 Data availability

Referring to the assumption made in the Method section, data refresh rate is kept equivalent on both data processing approaches. Regardless of the selected infrastructure, data availability comprises of two steps: data supply from EHRs and availability of extracts to statistical analyses. Ensuring reliable data collection from EHRs may be challenging due to varying network infrastructure, power outages and human factors. However, this step is equally challenging in both approaches, while the complexity of the next one depends on the selected architecture. Many techniques and tools exists for delivering high availability of a centralized system populated with data. It is, however, often complicated to reach high availability in a distributed infrastructure where data extracts are also distributed and every node needs to be highly available for data processing. On the other hand, in many cases data processing does not need to be performed at the exact moment, when the query is created. Tasks targeting certain nodes could be pooled until expected coverage is reached and results for the available network is computed. It could minimize data availability concerns in the distributed architecture scenario.

3.7 Cross-border coverage

Expanding the scope of the LHS across region/country borders or linking several national LHS is an important functionality in a long-term development of the health data reuse infrastructure. It may be more relevant for systems covering relatively small populations to maximize the number of cases, especially for rare conditions. Legal obstacles often arise when trying to create such knowledge bases.

It could be rather complicated or in some cases impossible to reach international coverage of the LHS using centralized data reuse infrastructure. Varying legal regulations regarding privacy preservation in patient data need to be addressed and finding a compromise could easily become a long and cumbersome process requiring changes in legal requirements in all participating countries. Storage of such information outside the country of origin is a known restriction without a clear solution.

Distributed LHS infrastructure may be seen as a solution addressing the aforementioned challenges. It assures, that patient data does not leave the country of origin, voiding many of the legal barriers. Regardless of the easier adaptation to the existing legal frameworks, many problems related to data interoperability, management of the infrastructure, sustainable funding and other aspects need to be addressed.

4 DISCUSSION

The comparative analysis summarized strong and weak points of choosing distributed or centralized architecture for implementing the LHS. A brief list of summary points is represented in Table 1.

Table 1 provides argumentation why making a choice between the two approaches is a complex task, as it was indicated in a recent review on the implementations of the LHS (Budrionis and Bellika, 2016). Both approaches have their advantages and limitations, which could have different weights depending on the legal framework the system is implemented in, experiences with health data reuse and other factors. Even though ensuring the privacy of patient data should be the highest priority, it also minimizes the utility of data and could even lead to misleading results (Tucker et al., 2016). It calls for research and development of novel techniques for privacy preserving data processing.

Criteria	Distributed	Centralized
1. Data access and timeliness	+	+*
2. Capacity of statistical analyses, trust in results	-	+
3. Control of data	+	-
4. Patient privacy/data security	+	-
5. Data quality and utility	+	-
6. Data availability	-	+
7. Cross-border coverage	+	-
Sum:	5+, 2-	3+, 4-

Table 1 Comparison of distributed and centralized data processing architectures for LHS (+ indicates that the infrastructure is sufficient for addressing the needs of the LHS, - lacks support to meet the requirements of LHS, * marks cases dependent on assumption of equivalent data refresh rate made in the Method section)

Assumptions made in Method section ensured that narrow purpose health registries were not included in the comparison due to their different function, which does not fit into the agile nature of the LHS. Instead, a multipurpose data storage was used in the comparison, meeting the requirements for establishing the LHS and highlighting the implications of centralized and distributed data processing. A term multipurpose in this context represents the potential utility of data to deliver benefits to various stakeholders. Such benefits should not be solely focused on clinicians in return to their data contribution or governmental bodies financing the initiative. The aforementioned assumptions made a centralized data storage a comparable counterparty, however, it is not yet clear how such information resource should be built and

maintained. There are many unanswered questions in large-scale distributed architecture as well, however, in a long run it has the potential to address all anticipated data reuse needs and overcome the bottlenecks of the centralized systems (Brown et al., 2010).

This paper has not looked into the potential of "hybrid approaches" combining the strengths of both distributed and centralized architectures. Centralization could be performed at various levels (for instance, commune), while keeping the advantages of distributed architectures in higher levels. Such choice would potentially increase the complexity of the overall system, but could contribute to addressing the weaknesses of both architecture choices summarized in Table 1. This approach calls for more research evaluating its feasibility and potential benefits.

5 CONCLUSIONS

This paper looked into the advantages and limitations of distributed and centralized health data processing infrastructures to support the development of the LHS. It provided argumentation for making the decision between the two architectures. The conclusions of the comparison were consistent with the findings of a recent literature review (Budrionis and Bellika, 2016) and global development in the field. A straightforward clear-cut answer to the question "distributed or centralized?" does not seem to exist and many important aspects need to be considered in the decision process.

Several discussed criteria were in favor of the distributed data access and processing architecture as a fundament for the LHS. Capacity of the available statistical methods, trust in the results and data availability were regarded as major limitations (Table 1), however, even more weak points were identified in the centralized approach. It calls for research to address the deficiencies of both approaches to meet the requirements of the LHS infrastructure better.

6 REFERENCES

[1] Asfaw Hailemichæl, M., Yitbarek Yigzaw, K., Bellika, J.G., 2015. Emnet: a System for Privacy-Preserving Statistical Computing on Distributed Health Data, in: Proceedings from The 13th Scandinavien Conference on Health Informatics. Presented at the The 13th Scandinavien Conference on Health Informatics, Linköping University Electronic Press, Linköpings universitet, Tromsø, Norway.

[2] Bellika, J.G., Henriksen, T.S., Yigzaw, K.Y., 2015. The snow system: A decentralized medical data processing system. Methods Mol. Biol. Clifton NJ 1246, 109–122. doi:10.1007/978-1-4939-1985-7_7

[3] Brown, J.S., Holmes, J.H., Shah, K., Hall, K., Lazarus, R., Platt, R., 2010. Distributed Health Data Networks: A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care. Med. Care 48, S45. doi:10.1097/MLR.0b013e3181d9919f

[4] Budrionis, A., Bellika, J.G., 2016. The Learning Healthcare System: Where are we now? A systematic review. J. Biomed. Inform. 64, 87–92. doi:10.1016/j.jbi.2016.09.018

[5] Greenhalgh, T., Howick, J., Maskrey, N., 2014. Evidence based medicine: a movement in crisis? The BMJ 348, g3725. doi:10.1136/bmj.g3725

[6] Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007. The Learning Healthcare System: Workshop Summary, The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC).

[7] Morris, Z.S., Wooding, S., Grant, J., 2011. The answer is 17 years, what is the question: understanding time lags in translational research. J. R. Soc. Med. 104, 510–520. doi:10.1258/jrsm.2011.110180

[8] Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M.J., Williams, Z., 2016. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med. Res. Methodol. 16. doi:10.1186/s12874-016-0169-4

[9] U.S. Department of Health and Human Services, Office for Civil Rights, n.d. Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. Breaches Affecting 500 or More Individuals [WWW Document]. URL https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (accessed 5.16.17).

[10] Yigzaw, K.Y., Bellika, J.G., Andersen, A., Hartvigsen, G., Fernandez-Llatas, C., 2013. Towards privacy-preserving computing on distributed electronic health record data. ACM Press, pp. 1–6. doi:10.1145/2541534.2541593

7 ACKNOWLEDGEMENT

This research was funded by a grant from the Research Council of Norway to the Norwegian Centre for E-health Research, University Hospital of North Norway. Grant number 248150/O7