






# Empowering Translational Health Data Science Capabilities in Population Health Management

## A Case of Building a Data Competence Center

Armel Lefebvre<sup>1,2</sup> , Lisette de Schipper<sup>1</sup>, Marcel Haas<sup>1</sup> ,  
and Marco Spruit<sup>1,2</sup> 

<sup>1</sup> Department of Public Health and Primary Care, Leiden University Medical Center,  
Albinusdreef 2, 2333ZA Leiden, The Netherlands

a.e.j.l.lefebvre@lumc.nl

<sup>2</sup> LIACS, Leiden University, P.O. Box 9512, 2300RA Leiden, The Netherlands

**Abstract.** In this paper we present the outcomes of a survey conducted among the research community of a Population Health Management department. The goal is to investigate how (translational) data science applications can be supported in a complex ecosystem of data sources and regulations of secondary healthcare data use. The envisioned solution is the creation of a data competence center as a multidisciplinary unit mixing research and professional support staff to provide data science technology, training, and resources to (early-career) researchers to address current challenges that are considerably impacting data quality and reproducibility in PHM research.

**Keywords:** Population Health Management · Data Infrastructure · Data Competence Center · Translational Data Science · Reproducibility

## 1 Introduction

Population health management aims at improving the fairness and efficiency of prevention and intervention programs in healthcare. First, there are opportunities to reduce costs in healthcare with prevention tailored to populations at risk. Furthermore, intervention programs may alleviate the need for people to undergo medical treatment if specially designed intervention tools effectively reduce (future) health risks while reducing costs growth of healthcare [1]. The primary source of evidence for PHM are data from electronic health records (EHR) [2], which are used to obtain individual measurements from general practitioner (GP) practices and hospitals. EHRs contain both structured (e.g., disease identifiers) and unstructured (e.g., additional observations by the GP during consultation) data.

To exploit EHR data for research purposes, agreements are made between research performing organizations (e.g., and (local) GP practices to obtain patient records in a systematic way. This is also the case for the PHM data infrastructure we are investigating in this paper, which is named ELAN and stands for Extramural Leiden University

Medical Center Academic Network [3]. In ELAN, agreements have been made with GP practices in the South Holland region, and other EHR data holders such as regional hospitals.

Thus, the construction of ELAN involves organic growth of disparate data sources as well as individual agreements with third parties to provide data for research and increase regional coverage. This is a situation that echoes other PHM initiatives making use of EHR data under a meaningful use scheme [2], where (privacy) regulations involve strict data access rules. This involves data quality and data integration challenges before researchers in PHM can analyze data for research purposes. This complexity has yet to be reduced for PHM researcher to make the research process in PHM itself more effective.

To achieve that, we are introducing a specialized unit in the ELAN infrastructure, called a data competence center, to tackle those issues and facilitate the work of researchers. This is particularly relevant when the demand for EHR analysis is relying more often than not on data science applications that require high-quality data and smooth data access workflows. Hence, the main research question (MRQ) of our study is as follows:

**MRQ: “How can a data competence center (DCC) strengthen translational data science capabilities in a Population Health Management context?”**

Consequently, we will seek to define the requirements expressed by researchers of PHM infrastructure (PHI). Those requirements were collected to identify how PHM researchers can be better supported during their research.

In the remainder of this paper, we first introduce the background of our study, especially the more niche areas of population health management and translational health data science. Then we briefly explain how we collected evidence from researchers in population health management who are employing data science approaches to perform their research. Finally, we conclude this paper with an agenda for building data science capabilities in a PHM context.

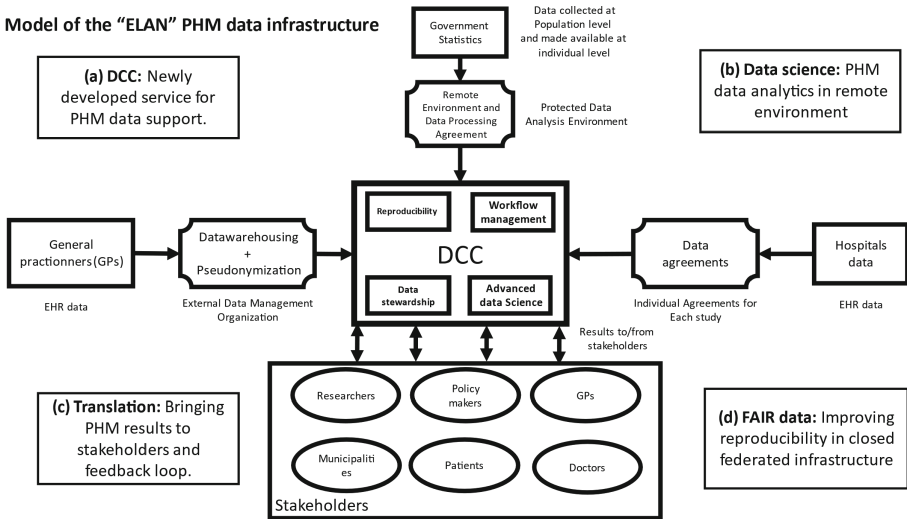
## 2 Background

### 2.1 Population Health Management (PHM) Infrastructure

PHM studies combine health data with socio-economical insights about individuals or (regional) populations. Therefore, a robust data infrastructure is crucial [4]. In a PHM context, this data infrastructure is based on federating sources from independent hospitals, networks of general practitioners, and databases of governmental statistics. In case of the Extramural Leiden Academic Network (ELAN) infrastructure, nine data sources can be combined to provide data for researchers [3]. Besides, external data sources, such as those named earlier, are often also integrated with data collected from researchers during field or survey studies.

As can be seen from Fig. 1, in ELAN there are three types of data providers which are GP data, Governmental organizations and (regional) hospitals. Agreements with GPs enable the gathering of data in a secure data warehouse, which is then queried according to the specific needs of a research project. Other type of data providers,

such as hospitals or governmental organizations will refine data access authorizations to individual researchers. This involves administrative steps to access health data at individual level at the start of a research project.



**Fig. 1.** The role of a DCC (a) DCC is the new service described in this work, it is embedded in the PHM infrastructure and collaborates with the three ELAN managerial units, (b) **Data science** applications take place in a secure environment, the technical skills and tools discussed in this paper relate to this environment, (c) **Translation** is the act of deploying results to the community of stakeholders as well as obtaining feedback on the results efficiently, and (d) **FAIR data** focuses on the role of the DCC to support FAIR data exchanges.

## 2.2 Digital Competence Center

The creation of a digital competence center (DCC) follows some more recent trends in the area of research data management [5, 6]. DCCs address a common burden of research projects where a specialized unit which is knowledgeable of the type of research conducted, the technology field to effectively deploy data analytics and support data management that adhere to FAIR principles for both data and software [7, 8]. For ELAN, the DCC is composed of PHM researchers, data scientists and data managers who are focusing on making PHM research more reproducible, seek to make data analysis workflows more efficient and train researchers in more advanced data science skills (as explained hereunder in Sect. 4).

## 2.3 Translational Data Science

Currently, ELAN data science activities are performed by researchers in a secure environment as depicted in the upper-right box (Fig. 1.b). The further goal of establishing a

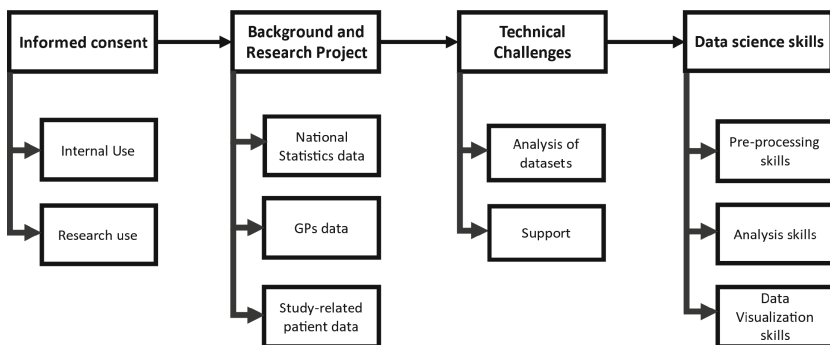
DCC is to help PHM researcher embrace more advanced technology and uses of data science to broaden the insights from EHR data. Translational data science (TDS - Fig. 1.C) is a field of research dedicated to bridging data science with efficient application of data analytics in clinical and healthcare settings [9–11]. As we will discuss later in this paper, bringing data science skills to a level where TDS can be successfully enacted has its challenges, which the DCC and its FAIR data management ambition described in Fig. 2.D aims to tackle.

### 3 Method

We investigated the current state of the PHM infrastructure using an exploratory single case study [12]. We started with the high-level requirements for implementing advanced data science capabilities in ELAN that were discussed by the management board in September 2023. Then, we opted for a survey to collect evidence from the (local) research community on the current status of data science skills and requirements.

The survey focused on ELAN users and support professionals at the department of Population Health Management, at the health Campus in The Hague in the Netherlands, which is part of the Public Health and Primary Care department of the Leiden University Medical Center. In September 2023, the management team presented several steps to be taken to professionalize the ELAN from a governance and infrastructure point of view. In the new governance framework, the ELAN is delegating some operational tasks related to data management to the ELAN-DCC. Among those operational (support) tasks, the ELAN-DCC ought to facilitate advanced data science methods on ELAN data.

## PHM Infrastructure Survey Workflow



**Fig. 2.** PHM Survey workflow. The survey is divided into three main sections: Research background, Technical Challenges and Data science skills.

In this context, we conducted a survey to obtain some more detailed insights on the data science skills and requirements of the users of ELAN. The main purpose is to identify the areas where the (ELAN) DCC can further enhance the support for data science activities for (early-career) researchers.

The ELAN-DCC survey opened at the end of February 2024 and closed on March 15, 2024. There were 27 respondents who completed the survey. We shared the survey link by e-mail among 95 potential respondents listed as ELAN users by ELAN data managers. We have effectively shared the survey link with 88 potential respondents (i.e., excluding emails bouncing back and out of office notifications). We obtained 27 responses, which is equivalent to a response rate of 31% (27/88). From those 27 respondents, one individual opted out, they did not allow the use of their data for research purposes. Therefore, we are examining the responses of 26 participants for the remainder of this paper.

The survey data was collected using a Microsoft Office365 online form. The form was divided into several subcategories (see Fig. 2) that could be shown or hidden depending on the kind of data sources used by the respondents. For a subset of respondents, we also obtained their e-mail addresses for future use. The data was downloaded as an excel file and processed in OpenRefine to standardize free text answers, correct the syntax of column names, and anonymize results for further processing. After processing the data with OpenRefine, the resulting dataset was uploaded in Python, using the Pandas module. The analyses and plotting were performed in a Python Jupyter notebook using the Pandas and Seaborn modules.

## 4 Results

In this Section, we present the main results that correspond to the survey categories depicted in Fig. 2.

### 4.1 Research Background

We note that the survey yielded answers from a majority of early career researchers who were appealed by the fact that more support for data science is being implemented at ELAN. This is reflected by the number of PhD candidates and Researchers/postdocs who are included in the sample (as described in Table 1).

**Table 1.** Number of respondents per academic role. Classified by seniority (Full professor, most senior) to PhD candidate (most junior role).

Academic role	Number of respondents
Full professor	1
Associate professor	2
Assistant professor	2
Researcher/postdoc	5
PhD candidate	16
Total	26

Some more senior academics are involved with supervision of early-career researchers, which brought some more knowledge about trends in PHM on the longer

term whereas early-career researchers could report on more direct burdens encountered with ELAN data in daily research.

## 4.2 Technical Challenges

For operations on the data, there are several aspects revealed in the free text answers to the question “What kind of analyses and operations do you consider to be unnecessarily complicated?” Recurring themes are challenges in data and file management, documentation, and data quality as well as data cleaning, selection, and processing.

For data management, the lack of data standards, naming conventions and inconsistent file paths impeding proper versioning are adding complexity when transferring data. Then, there are data sources that are not properly documented and there is no clear provenance information explaining why data is missing in some instances. Those data integration issues were not all addressed by even more proficient ELAN users who are keen on programming data processing scripts.

## 4.3 Data Science Skills

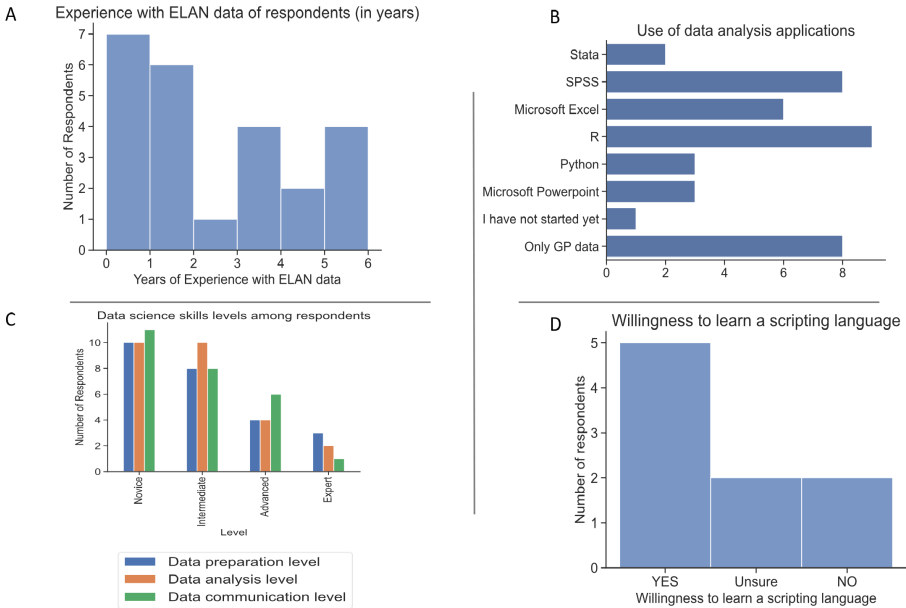
In Fig. 3 we present the main outcomes indicating that the areas in which the DCC will intervene have room for progress. In Fig. 3.A, we see that respondents have mostly one to two years of experience with ELAN data. Figure 3.B shows that while R is a popular scripting language, standard statistical packages and Excel files are frequently used.

Furthermore, Fig. 3.C shows variations in skills among respondents, while the majority is self-reporting at novice or intermediate level a point of attention is that the skill levels are not systematically a representation of a realistic skill level as experienced by users.

A follow-up question in our survey asked respondents “If you feel like adding more nuance/explanation to your choices, feel free to do so here.”, indicated that some respondents were somewhat in-between intermediate and advanced, or struggling with fundamental data integration tasks while being at an intermediate level on a scale of four levels (Novice, Intermediate, Advanced and Expert).

Finally, in Fig. 3.D showing the results on the willingness to learn more advanced programming skills, 37% of the respondents use a scripting language already. By scripting language, we include languages like R and Python, they are interpreted languages and have a vast ecosystem of packages to conduct (advanced) data analyses (see Table 2). Among the respondents who are not using scripting languages, the willingness to learn a scripting language is mixed.

Learning scripting languages has limitations that relate to the research environment. Respondents reported limited learning resources inside the organization to tackle more advanced analyses using programming scripts. Then there is a time issue, where early-career researchers communicated that the duration and requirements for completing a PhD trajectory left little room for learning programming skills. Last, data cleaning necessitates more examples when done with scripts, as there is little code exchange at the moment, repetitive cleaning tasks are coded from scratch.



**Fig. 3.** The distribution of experience, skills, and data science applications within the ELAN user community.

**Table 2.** Rationales of respondents to use scripting languages for their research.

Statements in favor of using scripting languages in PHM	Support
The types of analyses I perform cannot be performed in Excel/SPSS/STATA, so I am required to use a scripting language	4
I feel more comfortable performing research with a scripting language as opposed to software like Excel/SPSS/STATA	3
A data scientist does not use SPSS, for many reasons	1
I use a scripting language for transparency (Most scripting languages are open source and can be shared on platforms such as GitHub)	1
Last 2 bullets combined Transparency and reproducibility, and also more comfortable with (is the only thing I learned)	1

## 5 Discussion

In terms of requirements for data analytics we see two trends from the survey we collected. On the one hand, there is a trend that limits the maximal effort to invest in data analysis compared with the requirement to publish results on time (for instance during a PhD). This also contributes to researcher seeking advice or support in their immediate surroundings, whether by searching online or asking colleagues instead of spending more time finding resources inside the larger organization. On the other hand, there is

a concrete requirement for more advanced applications and the ease of data integration, analysis, and visualization in ELAN. Those results show that the creation of a multidisciplinary unit such as a DCC can be suitable solution to tackle those trends by reducing the efforts needed to find examples and training materials as well as become a known support option for researchers.

Still, a major limitation of the current state of our study is that we are limited by the input from users of ELAN. As part of future research, we aim at triangulating our local survey responses with respondents from more sites using similar healthcare data infrastructure in the Netherlands and abroad. Besides, we plan to conduct in-depth interviews to obtain more granular evidence on data management and analysis practices in different research projects and settings.

Furthermore, the survey is designed especially for the ELAN context. The combination of healthcare data from GPs and governmental data re-occurs in other Dutch initiatives such as AHON (University Medical Center Groningen) [13] and similar integrations occur internationally such as in the Healthdata platform (Sciensano, Belgium) [14]. In other words, our survey suffers from a limited generalizability due to our convenient sampling strategy. Also, the construct validity of several items needs further evaluation before collecting evidence from users of different health data infrastructure than ELAN to increase the clarity and generalizability of the survey questions to fit international health data initiatives.

That being said, while our study is limited to a single case study site, similar PHM infrastructure initiatives are created in the Netherlands. Our survey is the first step towards getting a better overview on the tools and skills involved in PHM and it can be deployed on any PHM infrastructure consisting of the statistical governmental data mixed with GP network data. The survey can therefore be considered as pilot material for further research for broader efforts including the European Health Data Space.

We have seen that ELAN users have a diversity of data analytics skills. In the meantime, data analytics skills have room for improvements across the complete set of ELAN users, which is an aspect to be tackled by the ELAN-DCC.

## 6 Conclusion

This study aimed at answering the question: “How can a data competence center (DCC) strengthen translational data science capabilities in a Population Health Management context?”. We have brought preliminary insights from practice that serve to delineate the shape and role of a DCC in a PHM context. We have observed that a DCC needs to tackle data management, advanced data science skills and reuse of code to offer a robust infrastructure for translational data science applications to researchers. Especially, the need for the availability of data analysis recipes and materials, as pre-processing steps are similar although not shared among projects. As explained previously, a data infrastructure like ELAN is built upon a complex ecosystem of data providers which is currently scaling up and facing stringent privacy and use of AI (in healthcare) regulations.

The current study outlined the basic blocks of a data competence center and its role in population health management as a top-down unit to be implemented in PHM infrastructures to enhance data stewardship and data management. Based on preliminary



requirements from researchers and support professionals in one PHM research department in the Netherlands, we have identified pillars of a robust PHM DCC to properly address pressing challenges related to the use of data science for better healthcare. Moreover, we also include elements that make DCC-ready for next generation data analyses in PHM. Due to its multidisciplinary nature, a DCC can play a pivotal role in building up data science capabilities.

**Disclosure of Interests.** The authors have no competing interests.

## References

1. Steenkamer, B.M., Drewes, H.W., Heijink, R., Baan, C.A., Struijs, J.N.: Defining population health management: a scoping review of the literature. *Popul. Health. Manag.* **20**, 74–85 (2017)
2. Kruse, C.S., Stein, A., Thomas, H., Kaur, H.: The use of electronic health records to support population health: a systematic review of the literature. *J. Med. Syst.* **42** (2018)
3. Ardesch, F.H., et al.: The introduction of a data-driven population health management approach in the Netherlands since 2019: The Extramural LUMC Academic Network data infrastructure. *Health Policy* **132**, 104769 (2023)
4. van Ede, A.F.T.M., Stein, K.V., Bruijnzeels, M.A.: Assembling a population health management maturity index using a Delphi method. *BMC Health Serv. Res.* **24** (2024)
5. Jäkel, R., Peukert, E., Nagel, W.E., Rahm, E.: ScaDS Dresden/Leipzig – a competence center for collaborative big data research. *it Inf. Technol.* **60**, 327–333 (2018)
6. Ringersma, J., Adamse, P.: Data Stewardship@ WUR: advice on a role for Data Stewards. Wageningen Data Competence Center (2019)
7. Lamprecht, A.-L., et al.: Towards FAIR principles for research software. *Data Sci.* **3**, 37–59 (2020)
8. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016)
9. McDavid, A., et al.: Eight practices for data management to enable team data science. *J. Clin. Transl. Sci.* **5**, e14 (2021)
10. Slade, E., et al.: Integrating data science into the translational science research spectrum: a substance use disorder case study. *J. Clin. Transl. Sci.* **5**, e29 (2021)
11. Spruit, M.: *Translational Data Science in Population Health*. Leiden University (2022)
12. Takahashi, A.R.W., Araujo, L.: Case study research: opening up research opportunities. *RAUSP Manag. J.* **55**, 100–111 (2020)
13. Twickler, R., et al.: Data resource profile: registry of electronic health records of general practices in the north of The Netherlands (AHON). *Int. J. Epidemiol.* **53** (2024)
14. European Observatory on Health, S., Policies, Minna, H.: *Towards the European health data space: from diversity to a common framework*. World Health Organization (2022)