Universiteit Leiden
The Netherlands

The introduction of a data-driven population health management approach in the Netherlands since 2019: the Extramural LUMC Academic Network data infrastructure

Ardesch, F.H.; Meulendijk, M.C.; Kist, J.M.; Vos, R.C.; Vos, H.M.M.; Kiefte-de Jong, J.C.; ... ; Struijs, J.N.

Citation
Ardesch, F. H., Meulendijk, M. C., Kist, J. M., Vos, R. C., Vos, H. M. M., Kiefte-de Jong, J. C., … Struijs, J. N. (2023). 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*. doi:10.1016/j.healthpol.2023.104769

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3643506

Note: To cite this publication please use the final published version (if applicable).

Health policy 132 (2023) 104769

Contents lists available at ScienceDirect

Health policy

journal homepage: www.elsevier.com/locate/healthpol

The introduction of a data-driven population health management approach in the Netherlands since 2019: The Extramural LUMC Academic Network data infrastructure

F.H. Ardesch [a,*], M.C. Meulendijk [a], J.M. Kist [a], R.C. Vos [a], H.M.M. Vos [a], J.C. Kiefte-de Jong [a], M Spruit [a,b], M.A. Bruijnzeels [a], M.J. Bussemaker [a,c], M.E. Numans [a], J.N. Struijs [a,d]

[a] Department of Public Health and Primary Care/Health Campus The Hague, Leiden University Medical Center, the Netherlands
[b] Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands
[c] Leiden Institute of Public Administration, Leiden University, the Netherlands
[d] National Institute for Public Health and the Environment, Bilthoven, the Netherlands

ARTICLE INFO

Keywords:
Data infrastructure
Population health (management)
Integrated care
Routine care data

ABSTRACT

Improving population health and reducing inequalities through better integrated health and social care services is high up on the agenda of policymakers internationally. In recent years, regional cross-domain partnerships have emerged in several countries, which aim to achieve better population health, quality of care and a reduction in the per capita costs. These cross-domain partnerships aim to have a strong data foundation and are committed to continuous learning in which data plays an essential role. This paper describes our approach towards the development of the regional integrative population-based data infrastructure Extramural LUMC (Leiden University Medical Center) Academic Network (ELAN), in which we linked routinely collected medical, social and public health data at the patient level from the greater The Hague and Leiden area. Furthermore, we discuss the methodological issues of routine care data and the lessons learned about privacy, legislation and reciprocities. The initiative presented in this paper is relevant for international researchers and policy-makers because a unique data infrastructure has been set up that contains data across different domains, providing insights into societal issues and scientific questions that are important for data driven population health management approaches.

1. Introduction

There is emerging consensus that, in order to improve the health of (sub)populations, we need to adopt strategies that are not limited to the medical perspective, but have a broader perspective and pay ample attention to social – and other non-medical determinants of health [1]. This includes stable housing, nutritious food, social and cultural capital, education, income support, gender, ethnic context and lifestyle [2,3]. Strategies to improve health and reduce health inequality must focus on integrating services across and within different sectors like public health, health care in- as well as outside hospitals, social support, housing and community services [4]. In recent years, collaboration of multisector partners, often referred to as population health management [5,6] have emerged in many countries aimed to achieve better population health, (experienced) quality of care and a reduction in the per capita costs (Triple Aim) [7]. Well-known examples are Gesundes Kinzigtal [8] in Germany and Accountable Health Communities in the U.S. [9]. In the Netherlands similar multisector initiatives emerged in which public health, healthcare, social care and community services aim to organize and integrate their services [10]. Increasingly often, within these initiatives alternative payment models like bundled payments, shared saving contracts and population based payment models are explored and experimented with on a small scale [11,12]. Within these multisector partnerships the use of data and related analytic tools and instruments are more seen as a lever to improve the health of their populations and reduce health inequalities. However, most tools include routine registry data from one single domain or even one provider whereas it has been recognized that to improve population health, health equity needs to become a priority in the health sector [13] and measures of social determinants of health must be integrated into health

* Corresponding author
E-mail address: f.h.ardesch@lumc.nl (F.H. Ardesch).

https://doi.org/10.1016/j.healthpol.2023.104769
Received 21 January 2022; Received in revised form 27 February 2023; Accepted 9 March 2023
Available online 15 March 2023
0168-8510/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

programs and services to address disparities [13]. Consequently, an integrative data infrastructure that integrates the various data sources via data linkages is considered a prerequisite for an effective implementation for these multisector partnerships and applying concepts of population health management [4]. Within the Scandinavian countries such integrated data-infrastructures have been established including personal, health, housing, demographic and economic data [14,15]. All these data are collected, processed and accessible in line with the General Data Protection Regulation (GDPR) [16]. Knowledge gained from such longitudinally linked observational datasets has demonstrated to lead to a better understanding of population health, risk factors, patterns of care and does underpin the impact of the employed policies [17]. These insights are used to realize an optimal segmentation of the population according to shared adverse health risks, needed in order to implement proactive interventions, which in turn will lead to an efficient allocation of resources and insights regarding the decision which investments are worth the costs [6,18].

In this paper, we describe our approach towards the development of a regional integrative population-based data infrastructure. Through this data infrastructure, called Extramural LUMC Academic Network (ELAN), started in 2019, we aim to explore adverse health events and improve health of the population living in the greater The Hague and Leiden area. Initially, starting with routine primary care data, we focused on linking data from partners within the greater The Hague area and supporting their multisector collaborative partnership. This partnership resulted in a movement called Healthy and Happy The Hague (HHTH) and includes all relevant care professionals, social care providers, patient representatives and knowledge institutes. The format of bringing partners together and linking their data is now spreading across the region. We define (sub-) populations, identify care gaps, stratify according to shared risks, evaluate the employed proactive interdisciplinary interventions and programs and monitor whether the formulated key measures develop in the desired directions. A central premise of this regional partnership is that each partner agreed on sharing their own relevant (routine) data including electronic health record (ehr) data. In the remainder of this paper, we will discuss the implementation of these various data sources into a FAIR (Findable, Accessible, Interoperable, Reusable) centralized data infrastructure, describe how research results are disseminated through different types of deliverables to stakeholders, and discuss the challenges and opportunities of the ELAN data infrastructure. Simultaneously, we explore research opportunities in order to contribute to research methodologies in the context of population health management strategies.

2. The extramural LUMC academic network (ELAN) data infrastructure

The ELAN data infrastructure was set up as a centralized database infrastructure to which data sources supply through secure connections, aimed at data sharing and if required, coordination in the remote access environment of Statistics Netherlands (SN) . SN acts as the trusted third party and hosts the various data linked from external data sources (Fig. 1) [19]. All data suppliers and SN verified compliance with the General Data Protection Regulation (GDPR) before data linkage. Within the secured remote access environment of SN there is no possibility to de-pseudonymize the linkage key, which means that data can be analyzed on the individual level but only be reported at an aggregated level. As a consequence of this design of the data infrastructure, results can no longer be traced back to specific individuals. Therefore, translations of algorithms into tools for daily practice are subsequently needed.

All regional partners supply a part of the dataset. Currently, longitudinal data from Municipal Health Services (MHS), social support, mental health care, hospitals, acute and chronic care, primary care (i.e. GP practice centers) are available and updated regularly (Fig. 1). SN pseudonymizes all supplied individual data and removes identifiable information through a non-reversible process. After pseudonymization by SN, all data contain a so-called Record Identification Number (RIN) which makes it possible to link all data at the individual level within the ELAN data infrastructure. Subsequently, these datasets from external data providers can be linked with the data of the System of Social Statistical Datasets (SSD) [19]. The SSD covers integrated longitudinal microdata of numerous registers and surveys of the complete population of inhabitants of the Netherlands. Apart from the variables at individual

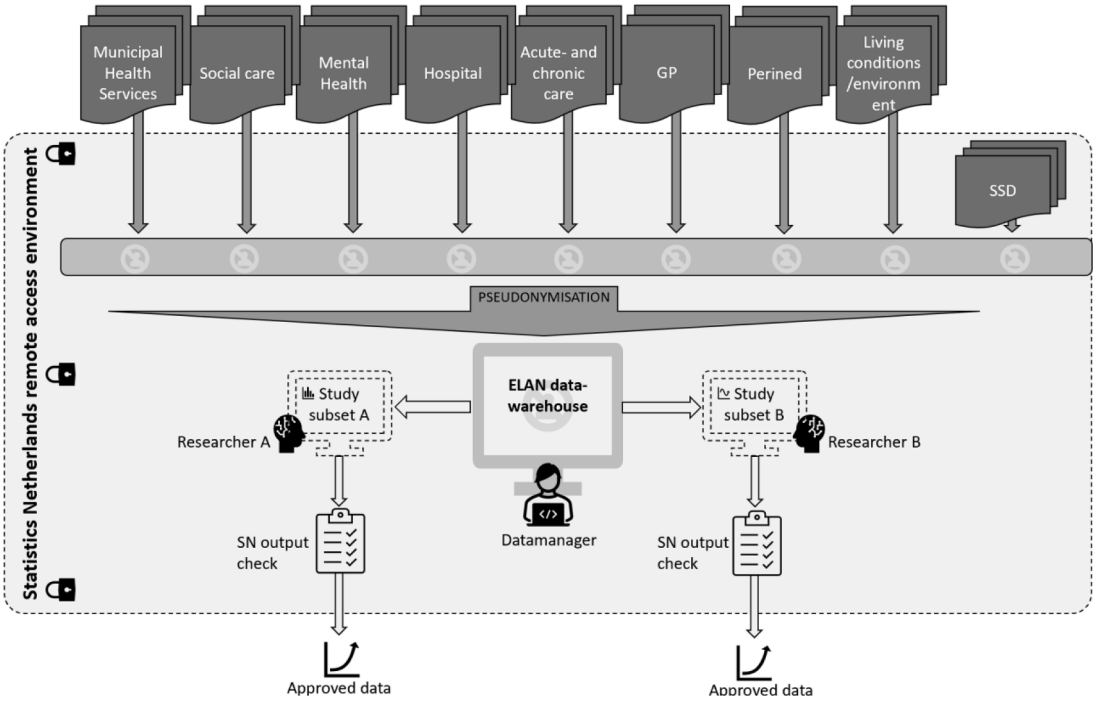

Fig. 1. Schematic overview of the ELAN data infrastructure.
Abbreviations: ELAN, Extramural LUMC Academic Network; GP, General Practitioner; SSD, Social Statistical Datasets, SN; Statistics Netherlands.

Table 1
Summary of the variables per data source, actor, years of availability and the number of included individuals (per February 2023).

Data source (Actor)	Variables	Available years of data	Number of individuals*
COVID-19 (Municipal Health Services-Haaglanden)	*COVID-19 positive lab tests, test date*	2020	6k
Social care (Municipality of The Hague)	*Provided services, request date, corresponding expenditures*	2011–2022	103k
Mental health care (Parnassia Groep)	*Coded Diagnosis (DSM-5), medication, health care utilization*	2011–2020	115k
Hospitals (Haga and HMC)	*Outpatient consultations, inpatient admissions, coded medication use (ATC) and diagnoses (ICD-10), lab results, COVID-19 outcomes*	2007–2022	300k (HMC) – 150k (Haga)
Acute care and chronic care** (Hadoks)	*Performed activities, ICPC, triage code, referrals, integrated care programs for Diabetes, COPD and Cardiovascular risk management*	2007–2022	Unknown
General Practitioners (ELAN)	*Consultations, coded medication use (ATC), diagnoses (ICPC), lab results and other coded physical measurements, contraindications and referrals*	2007–2022	800k
Perinatal data (Perined)	*Date of birth, birth weight and length, infant mortality, congenital abnormalities, adverse outcomes, characteristics of the (future) mother*	2007–2021	Nationwide
Living conditions/ environment (4-digit postal code level)	*Fine particles (PM10 to PM2.5), nitrogen, temperature/summer heat, noise pollution, green in the area*	2013–2021	Not applicable
SSD (SN)	Demographics: *gender, date of birth, ethnicity, marital status, job characteristics, education level, household composition*	2007–2021	Nationwide
	Income, debts and wealth: *individual/ household income, household debts and wealth, household savings*	2011–2020	Nationwide
	Neighborhood characteristics: *neighborhood and municipality codes*	2007–2021	Nationwide
	Mortality: *date of death, primary cause of death (ICD-10)*	2006–2021	Nationwide
	Aggregated yearly use of medication (ATC)	2007–2021	Nationwide
	Health monitor (survey): *lifestyle, perceived health, social environment*	2016 and 2020	Nationwide: 184k
	Medical specialist care: *healthcare activity, healthcare procedure and expenditure, specialism*	2013–2020	Nationwide
	The Long-term Care Act: *delivered care, care weight package use*	2015–2021	Nationwide
	Healthcare expenditures (limited to Health Insurance Act): *GP-care, pharmacy, oral care, hospital, paramedical, supporting medical tools, medical transport, birth care, care abroad, first line psychology, mental healthcare, geriatric care, home nursing, defaulters of the health insurers*	2011–2020	Nationwide
	Youth care: *youth care form, duration and perspective, characteristics of trajectories for juvenile probation, youth protection*	2015–2022	Nationwide
	Criminality: *registered victim support, registered detainees, registered criminal incidents*	2007–2021	Nationwide

Abbreviations: COVID-19: Coronavirus Disease 2019; DSM, Diagnostic and Statistical Manual of Mental Disorders; ATC, Anatomical Therapeutic Chemical; ICPC, International Classification of Primary Care; GP, General Practitioner; PM, Particulate Matter; SSD, social statistical datasets; HMC, Haaglanden Medisch Centrum; Hadoks, Haaglandse Dokters; ELAN, Extramuraal LUMC Academisch Netwerk; SN, Statistics Netherlands. * Refers to the number of individuals with data at some point between the stated period. ** Expected to be available in April 2023.

level, environmental data like the amount of greenery, fine particles and noise pollution are available at postal code level. Table 1 provides an overview of the current available data (February 2023) per data provider including the corresponding underlying variables and years of data collection. The de-identified datasets are loaded into a master database, accessible only to designated ELAN data managers. The data managers are then responsible for setting up a study subset for research purposes. SN reviews and agrees on all to be published outputs in order to guarantee that none of the published figures or tables can be traced back to identifiable individual persons or organizations (Fig. 1: output check).

To ensure access to data, information, and knowledge gained through the project, we followed the FAIR principles [20]. The FAIR principles are a set of heuristics designed to improve the infrastructures for data reusability and consist of four categories: Findability, Accessibility, Interoperability, and Reusability (FAIR). The very nature of the data infrastructure - especially its regular data updates and documentation - is aimed at ensuring reusability of the datasets that are brought together. Our modes of communication through websites, applications, dashboards, and social media ensure the findability of both the descriptions of the dataset and the acquired results. The different outputs are tailored for sharing with different users/user platforms - visualizations and interactive cross tables for policymakers, consumer-friendly applications for inhabitants, proportional datasets with researchers - we strive to meet the accessibility criterion. The well-documented and openly available data syntaxes and models used in the data infrastructure ensure the data's interoperability. Finally, the routine care data is regularly updated and can be reused by all researchers within the ELAN data infrastructure. A detailed summary of the applied FAIR principles and data governance is presented in Appendix B. In addition to the described linkage procedure, researchers can also decide to perform their analyses on the isolated (non-linked) dataset outside the SN remote access environment which has as advantage that it offers more possibilities for real-time innovative report tools for participating providers, but as drawback that potentially important variables are lacking in the analyses.

3. Potentials and objectives of the ELAN data infrastructure

The ELAN data infrastructure has multiple potentials and objectives to serve different actors. This section describes the value that ELAN can have at present, along with some examples. It is important to note that the examples provided are used as illustrative examples but not limited, and multiple projects and initiatives are explored or even undertaken.

3.1. Information for policy makers (both governmental and board members of care organization)

Acquiring representative information is indispensable for policy makers striving to develop data-driven proposals. The longitudinal timespan of the dataset makes it possible to address these issues by showing trends of changing socio-economic or health characteristics per neighborhood. As an application of the data infrastructure in the The Hague area, we made a publicly available online dashboard (www.gezondengelukkigdenhaag.nl/wijkprofielen) accessible to regional and national policy makers. This dashboard combines aggregated data in order to give accurate data visualizations of all neighborhoods of The Hague (Fig. 2). Central to this dashboard are the longitudinal, prospective key performance indicators (KPIs) which were defined in cooperation with policy makers affiliated with the HHTH-movement. The set of KPIs will be regularly reformulated based on the scientific literature [21] but also in order to address urgent societal and political questions concerning the region.

Another possibility that arises by using the periodically updated data infrastructure is that we are able to monitor the impact and effect of applied interventions without additional data collection. As we are able to combine data on health care utilization and outcomes, healthcare expenditures, and define comparable control groups, we are able to gain insights in potential effects of the interventions.

3.2. Providing relevant tools to health care professionals

Next to providing information on population needs to policy makers, we also construct risk stratification or predictive algorithms that provide relevant information for health care professionals. An example of such a tool was developed in the project of Heart for Women the Hague based on data from GPs, hospitals and the SSD, which aims to improve the prevention of cardiovascular events of women having had hypertension and/or diabetes during pregnancy [22]. An iterative process of adjusting care based on identification, developing risk prediction tools, process and outcome monitoring and evaluations was developed and implemented, together with patients, midwives, data scientists, medical specialists (gynecologist and internal medicine specialists), GPs and practice nurses. Similar tools were developed for structuring care for the elderly [23] and primary care patients carrying a risk for developing chronic somatic symptom disorder [24]. Essential for sustainability and long-term implementation of the regional data infrastructure as an infrastructural asset to be used in the health domain in the region, is that results of analyses can be fed back to healthcare professionals in a way that subpopulations with comparable risks can be recognized during daily practice. The current system makes results available on various levels, depending on the level of consent, the technical possibilities and privacy protective regulations. In the end it will allow healthcare providers in the region to bring population health management in practice.

3.3. Enhancing citizen involvement

To provide support for the implementation of policy measures and ensure social renewal by introducing ideas, subjects and approaches, citizen involvement is essential. We involve citizens by having active dialogues between decision makers and inhabitants, sharing study results, and identifying challenges towards implementation. To facilitate this, we cooperated for instance with action researchers working in the disadvantaged neighborhoods in The Hague. For example, within the neighborhood Moerwijk, we explored in cocreation with citizens which initiatives can be used to maintain or increase the health of citizens and reduce inequalities. The ELAN data infrastructure was used to ascertain which problems existed in the neighbourhood across the social and medical domain to provide guidance for the action researchers to combine quantitative data with the experiences and narratives of the residents [25]. The action researchers presented and discussed the results during 'neighborhood dialog meetings'. During these meetings, residents and health care professionals could prioritize the problems to tackle, and moreover, how to choose interventions that potentially have impact on citizen's daily lives. For example, as part of the 'Promising

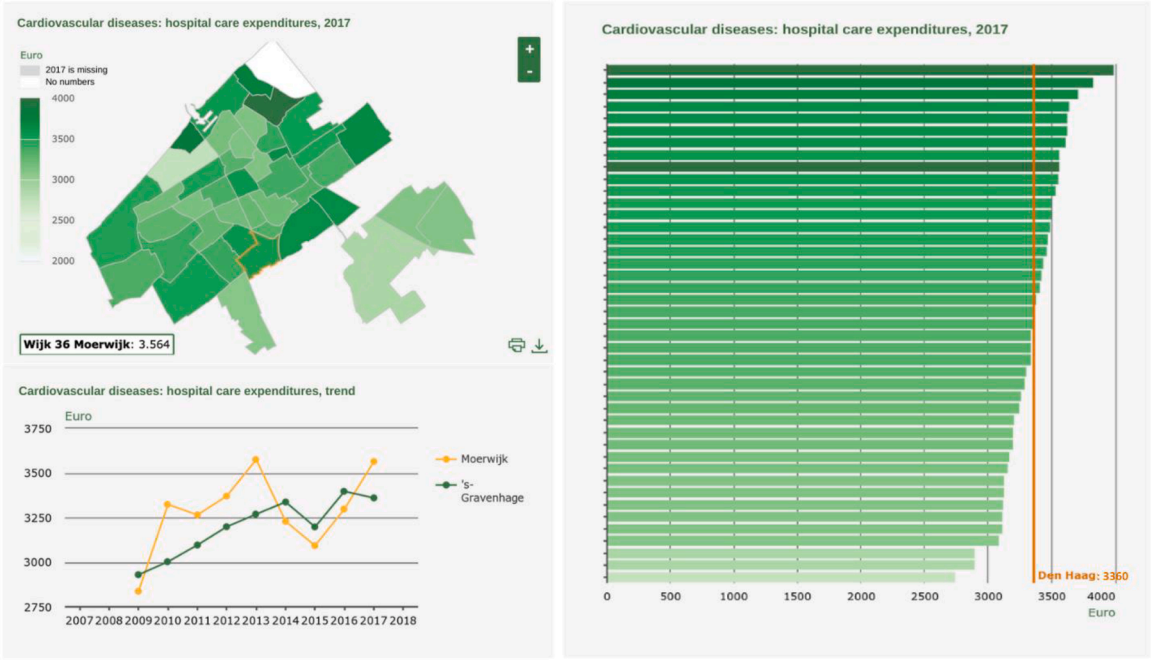

Fig. 2. Example of a neighborhood profile of The Hague.

Start' program activities [26] we identified that the first priority for single mothers with multiple problems was not to get more care or social support, but to have access to meeting places that provide room for social connectedness with other single mothers.

3.4. Increasing knowledge via research

With the ELAN data infrastructure, we aim to make data available for reuse in studies by researchers in various domains related to health. To obtain access for research purposes, researchers need to complete and submit an application form for a new research project. Their application is subsequently taken into consideration by the compliance committees of the data providers affiliated with the data infrastructure and when indicated with SN. When granted, a subset of data is made available to them by the data managers through the secure cloud environment, which is hosted by a SN (Fig. 1). The following organizations may be granted access to the secured remote access environment of SN and consequently to the ELAN subset of linked data: Dutch universities, institutes for scientific research, organizations for policy advice or policy analysis, statistical authorities in other EU countries and other research institutions authorized to work with the microdata. To date, the first scientific papers have been published based on the use of ELAN [22, 27–29] and, tailored to the needs of the stakeholders, we expect many more to come in the coming years as multiple researchers are working on scientific papers using ELAN data.

4. Early and future challenges and opportunities

Although the ELAN data infrastructure already has a wide variety of data sources available, ELAN currently lacks data that report patients' experiences (i.e. Patient Reported Experience Measures (PREMs) and Patient Reported Outcome Measures (PROMs)). A first attempt to add self-reported measures has been undertaken by developing the COVID radar app during the COVID-19 pandemic [30]. This app aims to measure self-reported COVID-19 symptoms, social-distancing behavior and self-reported COVID-19 infections. Also existing unsolicited patient ratings will be used to gain insight into provider-based patient experiences [31]. These self-reported data will be linked to our data infrastructure, thus enriching the ongoing studies. Secondly, we also linked data from the Dutch 'health monitor' executed by the National Institute for Public Health and the Environment together with all local Municipal Health Services. The 2016 health monitor is a nationwide questionnaire survey with more than 450 K respondents (which resulted in $n = 15.000$ in the area of The Hague) that provides insight into the health, social situation and lifestyle of the Dutch population. The Health Monitor is executed every four years [32].

Another limitation concerns the straight forward availability of uncoded, free text data to be extracted from primary and secondary care EHR. Privacy reasons may hamper access and accordingly reuse for text mining and other advanced techniques. Previous research has already shown that free text from GP data can be a valuable in addition to coded diagnoses [33,34] and to complement for missing coded data. For instance, a study of Groenhof et al. showed that approximately 90% of the smoking status can be found in the journal of the GP-information system when applying text learning techniques [35] and in another study cancer status could be corrected to over 90% by searching EMR texts [36]. Further possibilities to access unstructured medical record texts will be explored through improved text mining strategies in routine healthcare data.

The ELAN data infrastructure contains a wide variety of routine (care) data that can be linked within the SN remote access environment. Apart from the fact that analyses can only be performed on a certain aggregated level, another disadvantage is that some of the linked variables, i.e. income, although potentially being an important predictor, are unavailable in most EHR (Electronic Health Records). So, when the advice is to use low income as a predictor in an algorithm in daily healthcare, the relevant healthcare authority cannot trace the right individuals due to a lack of data and proxy-parameters have to be identified. Furthermore, privacy concerns continue to play a major role in linking data sources. We experienced that data source holders are reluctant to upload their data due to privacy related issues, but additionally are also insecure due to a (perceived) lack of control over their data once uploaded. To tackle both issues in the future, federated learning methodologies could be a promising option. Federated learning is a privacy-friendly form of machine learning which brings the machine learning model to the data instead of the other way around [37]. However, it should be noted that federated learning is new and not widely adopted yet and consequently has not fulfilled its expectations. It will take some time before federated learning techniques can be implemented widely within the ELAN data infrastructure.

Finally, working with routine care data always entails certain implications. Despite the fact that most of the data have already been standardized, we still have to take into account missing data, selection bias, linkage errors and administrative delay and other common issues [38]. However, when compared to regular observational studies, the SSD gives us the opportunity to include almost 100% of the Dutch population and thereby reduce selection bias and the percentage of data missing. In order to offer researchers as much support as possible, plausibility checks are performed before every upload of a data supplier (Appendix A). Methodological issues regarding the use of the data and advice to solve these issues within ELAN are described at a central point and made accessible for all researchers. In addition, regular ELAN meetings are organized in which data issues, linkage procedures and other methodological issues, but also innovative research ideas are discussed among researchers in order to create a common ground and support researchers as much as possible.

5. Lessons learned

In this section, we describe the four most important lessons we have learned so far in order to support similar future initiatives. The first lesson and also the first step in the entire process is to gain and maintain support at the executive level. When the executive level of the various actors is convinced of the benefits of sharing data, it is no longer a question if, but particularly how data will be shared. Secondly, we have noticed that data-sharing organizations find privacy and legislation to be a complicated matter. Legislation regarding data sharing sometimes contains gray areas and thus legal experts from different organizations have different views on the same legislation regarding data-sharing. To overcome this complex matter, we used the 'whole system in a room approach'. We brought several parties multiple times together and discussed their needs and concerns until a solution was found. Thirdly, we have learned that carefulness in the process takes precedence over speed. Initially, we were eager to make progress quickly, but stakeholders need time to implement cultural changes within their organization. Maintaining trust is crucial, as any mistakes in the process can disrupt the entire process. Additionally, organizations need time to implement changes, as data sharing can be a time-consuming and demanding task. Lastly, we experienced that reciprocity to the partners that provide data is important. The establishment of reciprocity plays a role in fostering trust between the partners and ELAN, as it evidences that the utilization of their data is consistent with their priorities and interests and thereby increasing the likelihood of their continued participation in data-sharing initiatives in the nearby future.

6. Conclusion and the way forward

In this paper we describe our attempt to build a regional data-infrastructure including data from both medical and social domain, started in 2019. By doing so, we aim to enhance the use of a data driven approach by all relevant stakeholders to rationalize their decision making, allocate their resources efficiently and ultimately contribute to

their formulated goals, while at the same time aim to further mature the research field of population health management. As we constructed a data infrastructure through routinely collected observational data, we reduce the administrative burden for data collection, while simultaneously increase data availability for researchers and reduce the cost of data collection. As the ELAN data infrastructure is increasingly fulfilling its promise, we explore the possibilities to extend its geographical reach for hospital, General Practice and mental health care data to require a higher coverage rate within the ELAN region and similarly look for new data from for instance nursing homes and youth healthcare. By extending the reach of ELAN, we provide our full network of partners with appropriate deliverables, our goal is to disseminate our research results in order to ensure their impact on population health and well-being. Through this approach, we hope to lead by example and improve population health management research methodologies. As population health management continues to take root in multiple countries, learning from how to link, reuse data and support decision making within primary care process and their successes and failures will be critical. Building a data-infrastructure like ELAN in order to improve population health and reduce inequalities requires building trust, overcoming cultural differences, and address legal and regulatory barriers.

Declaration of Competing Interest

All authors declare that they have no conflict of interest.

Financial support

Funded by the Leiden University Medical Center.

Acknowledgements

We gratefully acknowledge the contribution of the partners of ELAN and D.O. Mook-Kanamori for critically reviewing an earlier version of the manuscript.

Appendix A: process of data linkage

Unique pseudonymized linkage ID

After the data is uploaded by the data providers, data is pseudonymized after it has been sent to Statistics Netherlands (SN). A pseudonymization can be completed in two ways: pseudonymization of persons is done on the basis of Citizen Service Number (Dutch abbreviation: BSN) or linkage key (date of birth, gender, postal code (including house number) and year of validity of the postal code). The result of the pseudonymization is that the personal data is replaced by the so called Record Identification Number (RIN) which is an unique ID within the secured environment of SN. The RIN can be used to link all data within the ELAN data infrastructure. Once pseudonymized by SN, data will never leave the secured environment in its original or pseudonymized form due to safety reasons. The incoming data are pseudonymised within two weeks after uploading it to SN. Afterwards, SN destroys the supplied dataset after encryption.

Matching and accuracy of the linkage

After pseudonymization of an external dataset, SN provides a document which gives information about the percentage of matched individuals. This percentage is calculated by matching the BSN or the linkage key with the Dutch population register which contains all the (unique BSN and demographic characteristics of every citizen who are registered in a Dutch municipality. The BSN and the linkage key match around 99%−100% and 94%−99% cases respectively. It should be mentioned that the matching accuracy depends to a large extent on how accurately health care professionals register. For instance whenever the address contains a spelling error, the concerning cases will not match.

Standardization before linkage

Data from the SSD is standardized by SN before it is made available to researchers [19]. All external data that has been uploaded to the ELAN database originates from the information system of the relevant organization. The data is stored in a standardized manner within this information system of the concerning health organization. However GPs and hospitals use different information systems to store their data. Each system registers the data in a slightly different way and the columns might have different names. For instance, gender can be displayed in different manners like ‘M’, ‘Male’ or ‘Man’ throughout the different GP-systems. To solve this problem for the GP-data, a trusted third party (TTP) named STIZON standardizes all data from different GP-information systems. In the case of the two hospitals, no TTP is involved.

Plausibility checks after linkage and pseudonymization

After the external data has been made available within the remote access environment of SN, the data manager performs a number of plausibility checks among which loss of number of patients, information, the inclusion criteria have been correctly implemented by the data manager of the concerning organization and whether the number of missing and impossible values can be called plausible. For example, we performed the following plausibility checks on the medication file for multiple years: amount of missing, number of rows including incorrect values (letters instead of values), illogical prescription dates (<1990 or >2023) and the number of individuals with diabetes medication (and compared with open source data). A detailed description of all applied plausibility checks is available on request.

Appendix B: FAIR principles and data governance

To ensure these parties’ access to data, information, and knowledge gained through the project, we followed FAIR principles when constructing the ELAN data infrastructure. FAIR principles are a set of heuristics designed to improve the infrastructures for data reusability and comprises the following concepts Findable, Accessible, Interoperable and Reusable (FAIR) [20].

Findable

Our modes of communication (through websites, applications, dashboards, and social media) ensure the findability of both the data and the acquired results. We are exploring options to make metadata, in the form of a codebook, available at any time. The codebook provides an overview of all available variables of the ELAN data infrastructure and is regularly updated to ensure that the metadata is up to date.

Accessible

As described, the incoming data of each data supplier is pseudonymised by Statistics Netherlands (SN) and made available within a secure environment to data managers of ELAN. The ELAN data managers clean up data if necessary and link it to the microdata using linkable data at personal, company and address level (see Appendix A). The linked files result in a project database from which subsets for specific studies can be made available to the researchers. Furthermore, through the different types of information we aim to share results with different parties - visualizations and interactive cross tables with policymakers, and consumer-friendly applications with inhabitants - we strive to meet the accessibility criterion.

Interoperable

The well-documented and openly available data models used in the data infrastructure ensures the data's interoperability. The ELA data infrastructure is designed in such a way that data can be safely transferred to the Remote Access environment of SN. Patient files from participating data providers are sent to SN. Data is encrypted via one-way pseudonymization. This means that all traceable data will be deleted, and that pseudonymized data cannot be de-pseudonymized. After pseudonymization, SN adds a unique key variable (Record Identification Number) which can be used to link all databases within ELAN with each other.

Reusable

The very nature of the data infrastructure - especially its frequent data updates and documentation - is aimed at ensuring reusability of the collected datasets. All data within the ELAN data infrastructure are stored in a secured SN-RA environment, whereby data can be reused at all times. The original subset issued to the relevant researcher is archived by the ELAN data manager. Subsequently, it is obligatory for the researcher in the context of Good Research Practice (GRP) to save all analysis steps. This indicates that all syntaxes, scripts, or similar files must be stored so that all steps within the analyzes (including recoding and exclusions) are reproducible by third parties. Data and syntaxes used for research must be kept for fifteen years for legal reasons. Archiving also takes place within the secured SN-RA environment.

Data security and privacy

Due to the fact that ELAN is responsible for bringing together data from multiple caregivers, ELAN data strives for the highest possible security of the data. Therefore, all storage sites and transmission routes are adequately secured. Before a researcher can gain access to the data of ELAN, the 'procedure for obtaining access' must first be completed. Once approved, a Virtual Private Network (VPN) connection must be set up between the researcher's computer and the SN-RA environment. Next, the researcher uses two-factor authentication that consists of logging in with a personal token and verification via SMS-code. Once logged in, researchers will only have access to pseudonymised data for which permission has been given by the initial data provider. The Governance structure of ELAN can be found on the website of the Health Campus (https://healthcampusdenhaag.nl/en/).

In case a researcher wants to use output (e.g. for a scientific publication or neighbourhood profiles), SN performs output checks that can guarantee that the concerned output cannot be traced back to individual persons or institutions for privacy reasons [39]. SN performs output checks such as:

- All tables and similar output contain at least 10 units (unweighted) as the basis for each cell or data point.
- All modelled output must have at least 10 degrees of freedom, where the number of degrees of freedom is equal to: number of observations - / - number of parameters - / - other model constraints.
- In all frequency tables and similar output, no cell may contain more than 90% of the total number of units in the row or column. This prevents certain variables in a table from defining a recognizable group. Although no individual entity can be recognized, confidentiality has been violated because the information is valid for virtually every member of the group and the group is identifiable as such.
- In all quantitative tables and similar data, the largest contributor to a cell cannot contribute more than 50% of the cell total.

References

[1] Kassler W, Tomoyasu N, Conway P. Beyond a traditional payer–CMS's role in improving population health. N Engl J Med 2015;8(2):109–11. https://doi.org/10.1056/NEJMp1406838. 372.

[2] Cockerham W, Hamby B, Oates G. The social determinants of chronic disease. Am J Prev Med 2017;52(1S1):S5–12. https://doi.org/10.1016/j.amepre.2016.09.010.

[3] Marmot M. Social determinants of health inequalities. Lancet 2005;19-25(9464): 1099–104. https://doi.org/10.1016/S0140-6736(05)71146-6. 365.

[4] Struijs J, Drewes H, Heijink R, Baan C. How to evaluate population management? Transforming the Care Continuum Alliance population health guide toward a broadly applicable analytical framework. Health Policy 2014:522–9. https://doi.org/10.1016/j.healthpol.2014.12.003.

[5] Woulfe J, Oliver T, Zahner S, Siemering K. Multisector partnerships in population health improvement. Prev Chronic Dis 2010;7(6):A119.

[6] Steenkamer B, Drewes H, Heijink R, Baan C, Struijs J. Defining population health management: a scoping review of the literature. Popul Health Manag 2017:74–85. https://doi.org/10.1089/pop.2015.0149.

[7] Berwick D, Nolan T, Whittington J. The triple aim: care, health, and cost. Health Aff 2008;27(3):759–69. https://doi.org/10.1377/hlthaff.27.3.759.

[8] Hildebrandt H, Hermann C, Knittel R, Richter-Reichelm M, Siegel A, Witzenrath W. Gesundes Kinzigtal Integrated Care: improving population health by a shared health gain approach and a shared savings contract. Int J Integr Care 2010;10: e046. https://doi.org/10.5334/ijic.539.

[9] Tirpirneni R, Vickery K, Ehlinger E. Accountable communities for health: moving from providing accountable care to creating health. Ann Fam Med 2015;13(4): 367–9. https://doi.org/10.1370/afm.1813.

[10] van Vooren N, Steenkamer B, Baan C, Drewes H. Transforming towards sustainable health and wellbeing systems: eight guiding principles based on the experiences of nine Dutch Population Health Management initiatives. Health Policy 2020;124(1): 37–43. https://doi.org/10.1016/j.healthpol.2019.11.003.

[11] Remers T, Wackers E, Dulmen S, Jeurissen P. Towards population-based payment models in a multiple-payer system: the case of the Netherlands. Health Policy (New York) 2022;126(11):1151–6. https://doi.org/10.1016/j.healthpol.2022.09.008.

[12] Hayen A, van den Berg M, Struijs J, Gert G. Dutch shared savings program targeted at primary care: reduced expenditures in its first year. Health Policy 2021;125(4): 489–94. https://doi.org/10.1016/j.healthpol.2021.01.013.

[13] Heisler M, Navathe A, DeSalvo K, Volpp K. The role of US health plans in identifying and addressing social determinants of health: rationale and recommendations. Popul Health Manag 2019:371–3. https://doi.org/10.1089/pop.2018.0173.

[14] Falkentoft A, Andersen J, Malik M, Selmer C, Gaede P, Staehr P, Ruwald A. Impact of socioeconomic position on initiation of SGLT-2 inhibitors or GLP-1 receptor agonists in patients with type 2 diabetes - a Danish nationwide observational study. Lancet Reg Health Eur 2022:14. https://doi.org/10.1016/j.lanepe.2022.10030. 100308.

[15] Laugesen K, Ludvigsson J, Schmidt M, Gissler M, Valdimarsdottir U, Lunde A, Sorgensen H. Nordic health registry-based research: a review of health care systems and key registries. Clin Epidemiol 2021;13:533–54. https://doi.org/10.2147/CLEP.S314959.

[16] union EP. General data protection regulation. Intersof consulting; 2018. Retrieved from, https://gdpr-info.eu/.

[17] Marmot M, Allen J, Bell R, Bloomer E, Goldbatt P. WHO European review of social determinants of health and the health divide. Lancet 2012;380(9846):1011–29. https://doi.org/10.1016/S0140-6736(12)61228-8. Epub 2012 Sep 8.

[18] Chong J, Lim K, Matchar D. Population segmentation based on healthcare needs: a systematic review. Syst Rev 2019;8(1):202. https://doi.org/10.1186/s13643-019-1105-6.

[19] Bakker B, van Rooijen J, van Toor L. The System of social statistical datasets of statistics Netherlands: an integral approach to the production of register-based social statistics pubmed. Stat J IAOS 2014;30(2014):411–24. https://doi.org/10.3233/SJI-140803.

[20] Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 15(3):160018. https://doi.org/10.1038/sdata.2016.18.

[21] Hendrikx R, Drewes H, Spreeuwenberg M, Ruwaard D, Struijs J, Baan C. Which Triple Aim related measures are being used to evaluate population management initiatives? An international comparative analysis. Health Policy 2016;120(5): 471–85. https://doi.org/10.1016/j.healthpol.2016.03.008.

[22] Kist J, Smit G, Mairuhu A, Struijs J, Vos R, van Peet P, Numans M. Large health disparities in cardiovascular death in men and women, by ethnicity and socioeconomic status in an urban based population cohort. EClinicalMedicine 2021;40:101120. https://doi.org/10.1016/j.eclinm.2021.101120.

[23] Drubbel I, de Wit N, Bleijenberg N, Eijkemans R, Schuurmans M, Numans M. Prediction of adverse health outcomes in older people using a frailty index based on routine primary care data. J Gerontol A Biol Sci Med Sci 2013;68(3):301–8. https://doi.org/10.1093/gerona/gls161.

[24] Girwar S, Fiocco M, Sutch S, Numans M, Bruijnzeels M. Assessment of the Adjusted Clinical Groups system in Dutch primary care using electronic health records: a retrospective cross-sectional study. BMC Health Serv Res 2021;21(1):217. https://doi.org/10.1186/s12913-021-06222-9.

[25] Van Eijk C, van der Vlegel-Brouwer W, Bussemaker J. Healthy and happy citizens: the opportunities and challenges of co-producing citizens' health and well-being in vulnerable neighborhoods. Adm Sci 2023;13(2):46. https://doi.org/10.3390/admsci13020046.

[26] Struijs J, Hargreaves D. Turning a crisis into a policy opportunity: lessons learned so far and next steps in the Dutch early years strategy. Lancet Child Adolesc Health 2019;3(2):66–8. https://doi.org/10.1016/S2352-4642(18)30384-5.
[27] Nieuwenhuijse E, Struijs J, Sutch S, Numans M, Vos R. Achieving diabetes treatment targets in people with registered mental illness is similar or improved compared with those without: analyses of linked observational datasets. Diabet Med 2022;39(6):e14835. https://doi.org/10.1111/dme.14835.
[28] Nieuwenhuijse E, van Hof T, Numans M, Struijs J, Vos R. Are social determinants of health associated with the development of early complications among young adults with type 2 diabetes? A population based study using linked databases. Prim Care Diabetes 2023. https://doi.org/10.1016/j.pcd.2023.01.002. S1751-9918(23)00003-7.
[29] Kist J, Vos R, Mairuhu A, Struijs J, van Peet P, Vos H, Groenwold R. SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study. eClinicalMedicine 2023:57. https://doi.org/10.1016/j.eclinm.2023.101862. 101862, ISSN 2589-5370.
[30] van Dijk W, Sadaah N, Numans M, Aardoom J, Bonten T, Brandjes M, Kiefte-de Jong J. COVID RADAR app: description and validation of population surveillance of symptoms and behavior in relation to COVID-19. PLoS ONE 2021;16(6):e0253566. https://doi.org/10.1371/journal.pone.0253566.
[31] Hendrikx R, Spreeuwenberg M, Drewes H, Struijs J, Ruwaard D, Baan C. Harvesting the wisdom of the crowd: using online ratings to explore care experiences in regions. BMC Health Serv Res 2018;18(1):801. https://doi.org/10.1186/s12913-018-3566-z.
[32] RIVM. Health monitor adults and elderly. National Institute for Public Health and Environment; 2020. Retrieved from, https://monitorgezondheid.nl/gezondheidsmonitor-volwassenen-en-ouderen.
[33] Tate A, Martin A, Ali A, Cassell J. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. BMJ Open 2011. https://doi.org/10.1136/bmjopen-2010-000025.
[34] Hoogendoorn M, Szolovits P, Moons L, Numans M. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016;69:53–61. https://doi.org/10.1016/j.artmed.2016.03.003.
[35] Groenhof T, Koers L, Blasse E, de Groot M, Grobbee D, Bots M, Haitjema S. Data mining information from electronic health records produced high yield and accuracy for current smoking status. J Clin Epidemiol 2020:100–6. https://doi.org/10.1016/j.jclinepi.2019.11.006.
[36] Sollie A, Roskam J, Sijmons R, Numans M, Helsper C. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. BMJ Open 2016;6(9):e012669. https://doi.org/10.1136/bmjopen-2016-012669.
[37] Sheller M, Edwards B, Reina G, Martin J, Pati S, Kotrotsou A, Bakas S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10(1):12598. https://doi.org/10.1038/s41598-020-69250-1.
[38] Hemkens L, Contopoulos-Ioannidis D, Loannidis J. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ 2016;188(8):E158–64. https://doi.org/10.1503/cmaj.150653.
[39] Brandt M, Franconi L, Guerke C, Hundepool A, Lucarelli M, Mol J, Welpton R. Guidelines for the checking of output based on microdata research. Academia; 2010. Retrieved from, https://www.academia.edu/38191675/ESSNet_SDC_-_Guidelines_for_the_checking.pdf.