AAPG2023	**Graph4Health**		PRCE
Coordinated by	Philippe CHONÉ	60 months	€
H.4 – Santé publique, santé et sociétés			

From Healthcare Accessibility to Health Outcomes

A statistical and machine learning approach to large-scale graphs

Consortium: CREST (CNRS UMR 9194), CASD, UMS-011 (INSERM), ESSEC

Summary table of persons involved in the project

Partner	Last name	First name	Current position	Role & responsibilities	Involvement (pers.month)
CREST	CHONÉ	Philippe	Professor	Coordinator Leads WP 0	30 p.m
CASD	GADOUCHE	Kamel	Director	Partner's scientific leader Inv. WP 1, 2	3 p.m
ESSEC	LECUÉ	Guillaume	PR	Partner's scientific leader Leads WP 3, Inv. WP 4	30 p.m
UMS-011 (INSERM)	KAB	Sofiane	Epidemiologist	Partner's scientific leader Leads WP 2 Inv. WP 1 Tasks 3.1	20 p.m
UMS-011 (INSERM)	GOURMELEN	Julie	Statistician	Co-leads WP 1 Inv. WP 2, Task 3.1, 1.4	20 p.m
CASD	LIU	Pengfei	Research engineer	Co-leads WP 1 Inv. WP 2 Tasks 3.1	20 p.m
CREST	KRAMARZ	Francis	Emeritus PR	Leads WP 4 Inv. Task 3.1,	20 p.m
CREST	KHALEGHI	Azadeh	Associate PR	Inv. Tasks 3.3, 3.4	12 p.m
UMS-011 (INSERM)	ZINS	Marie	Director	Inv. WP 1, 2 Inv Task 3.1	6 p.m
CREST	WILNER	Lionel	Affiliate	Inv. WP 3 4	36 p.m
CREST	DALALYAN	Arnak	PR	Inv. Tasks 3.3	6 p.m
CREST	CHOPIN	Nicolas	PR	Inv. Task 3.3	6 p.m
CREST	UHLENDORFF	Arne	PR	Inv. 3.1, WP 4	15 p.m
CREST	SCHMUTZ	Benoît	PR	Inv. 3.1, WP 4	9 p.m
CREST	LEVENEUR	Pauline	PhD student	Inv. 3.1, WP 4	12 p.m

No significant change in the full proposal compared to the pre-proposal

I Proposal's context, positioning, and objective(s)

a Objectives and scientific hypotheses

The project has four objectives: it intends to (i) build a platform that can manage massive health data and make them usable to researchers; (ii) use the tools of graph theory to describe the healthcare system in a systematic and quantitative way; (iii) develop new machine learning tools to understand the shape of the graphs and predict how this shape affects health outcomes; (iv) use these tools to study policies that contribute to the efficiency of the French healthcare system.

To do so, Graph4Health brings together expertise in economics, data science, epidemiology, and data management/big data analytics. The team has been authorized to access the French National Health Data System (Système National de Données de Santé, SNDS), which records all inpatient and outpatient claims for the entire population living in France. Specifically, we have been granted access to all records of consultations and other medical procedures, drug prescriptions, and hospital acute-care admissions (including surgical and non-surgical procedures, obstetrics). The data cover the years 2008 to 2018. To our knowledge, this is the first time the French data regulator (Commission nationale de l'informatique et des libertés)

grants access to exhaustive health data covering such a long period of time. In its recommendation to the CNIL, the scientific committee CESREES says: "It is precisely because it makes this kind of studies possible that the SNDS is of such scientific interest." Data on healthcare providers (Référentiel des professionnels de santé) are also available and will be merged with the main data source.

We intend to represent the healthcare system as a series of bipartite graphs. A bipartite graph is made of two types of nodes: patients and categories of healthcare providers (e.g., generalist doctors, specialist doctors, nurses, physiotherapists, etc.). Edges connect patients and providers. A patient and a provider are connected if they have met at least once during the current year. The projection of bipartite graphs on the set of providers will reveal information about how patients are shared between different doctors. A pair of doctors are connected in such a projected graph if they share at least one patient. Similarly, patients are connected if they have some providers in common. Graph connectivity is a defining feature of what we hereafter call "the healthcare system."

To understand the formation of these graphs and to use them to estimate the *causal impact of healthcare accessibility on utilization and health outcomes* (prescriptions, emergency hospital visits, mortality, etc.) are central to our contribution. We will examine whether certain local configurations (defined in terms of local endowments of physical or human capital or of hospital market structure) are better able to deliver positive outcomes to patients. Particular attention will be given to the geographic distribution of healthcare supply. We will build indicators of potential access at the local level, characterize potential low-density areas, the so-called "medical deserts", and quantify their effects on patients' outcomes.

To estimate the effect of the healthcare graphs on health outcomes, we will take advantage of exogenous changes in the local supply of healthcare such as hospital closures or the retirement (or death) of providers. Using the above-described projected graphs, we will have a clear view of how doctors share patients and refer them to specialists, and more generally of how patients circulate in the healthcare system. We will examine if and how patient sharing differs across geographic areas and varies from year to year. More generally, we will exploit the geographic and time variations of the graph to assess how the organization of the healthcare system affects its performance and ultimately the patients' well-being.

Our approach is novel in multiple dimensions: even though graphs have been extensively studied in recent years to understand social networks, bipartite graphs still need a full-fledge treatment. Furthermore, the spatio-temporal aspect of the graph calls for novel statistical and machine learning techniques.

b Position of the project as it relates to the state of the art

Most of the existing literature either does not observe the individual connections between patients and healthcare providers or take them as given. For instance, Finkelstein et al. 2016, 2021 analyse the sources of geographic variations in healthcare utilization and in mortality, disentangling supply-side and demand-side drivers. But they consider only elderly patients covered by Medicare and they do not observe the connections between patients and doctors as they form. They identify only the effect of places whereas we aim at understanding the effect of the structure of the healthcare graphs on health outcomes. The unpublished work of Badinski et al. 2022, the first study on U.S. data that seeks to take into account providers on top of patients and places, also takes the connections between patients and providers as given. The same is true for Agha, Ericson, et al. 2022, who analyze coordination among providers in Massachusetts and compute indicators of referral concentration.

Yet patient-provider connections are the results of individual decisions. Therefore, we will model them and at the same time assess their impact on health utilization and outcomes. We will estimate the patients' costs of traveling to visit a doctor and a hospital, allowing for observed and unobserved patients' and providers' characteristics. We will document the heterogeneity in doctors' practice-styles and assess its implications for health outcomes.

We will use the results to compute accessibility indicators. As noted by Lucas-Gabrielli et al. 2022, the existing indicators account only for the spatial dispersion of patients and physicians and ignore all unobserved factors that govern strategic choices (referral by another professional, physicians' reputation and labor supply behavior, patient inertia in choosing a physician after the patient has moved, etc.). One overarching goal of our project is to assess the extent to which accounting for unobserved heterogeneity and strategic decisions changes the characterization of medical deserts.

Most of the existing studies (including those cited above) consider sub-populations of patients, e.g., samples of Medicare beneficiaries in the U.S., and hence they observe only a fraction of each doctor's total patient panel. As a result, they are vulnerable to measurement errors when they compute statistics on the bipartite graphs. By contrast, our data provides an exhaustive view of the healthcare networks and of the activity of all healthcare providers, which is essential to avoid biases when modeling outcomes under unobserved patient or physician heterogeneity, see Abowd et al. 1999 and WP4.

The practical treatment of the data also requires some new techniques. Our project involves a total volume of about 20TB, which comes with many challenges, see, e.g., Kraus et al. 2018. The challenge is not only to host such massive database but to also allow researchers to find target data and analyse them easily. Many existing approaches try to apply big data analytic techniques to healthcare data, e.g., Pastorino et al. 2019, Dash et al. 2019, Jake Luo et al. 2016. However, most of them only focus on overcoming the volume and variety challenges. While some work address challenges concerning veracity Baldominos et al. 2017, a much needed metadata management system to address the discoverability and usability challenges is still lacking.

c Methodology and risk management

A steering committee will supervise the progress of the research work and the fulfillment of the tasks: Ph. Choné, K. Gadouche, G. Lecué and S. Kab . We plan to organize ANR meetings twice a year, gathering all the participants, to exchange on advances and results, to coordinate our research work and to foster new developments. Some of the network members (Ph. Choné, J. Gourmelen and S. Kab) are already collaborating on a related project based on the Constances cohort.

The project involves no legal risk with regards to the data regulator or the data producer. The data access agreement with Caisse National d'Assurance maladie (Convention d'accès aux données) has been signed in January 2023 and part of the data (covering years 2016 to 2018) has already been delivered. The data for years 2008 to 2015 are yet to be delivered. We plan to interact closely with the Direction de la statégie, des études et des statistiques of Caisse nationale de l'Assurance Maladie, regarding both the content of the data and the substance of the research questions.

One partner of Graph4Health, namely UMS 011, is specialized in epidemiology and includes medical doctors and pharmacists (Marie Zins and Sofiane Kab). Yet the majority of the team members are experts in econometrics, statistics, and big data analytics. Taking advantage of the UMS 011's network, we will extend our collaborations with medical doctors and public health researchers to better connect our work to the medical literature. These collaborations will greatly help us choose health outcomes and care quality indicators that are relevant to the purpose of the project and can be constructed from our claims data.

Work Package 0: Project Management

Leader: Ph. CHONÉ (CREST)

Deliverables: Two advancement reports per year.

Work Package 1: Data Preparation

Leader: P. LIU (CASD) and J. GOURMELEN (UMS-011)

In this work package, first, we will build a data platform that provides the following tools: (i) big data analytics and visualization tools to allow researchers to extract insight; (ii) geographical information system for location and distance calculation; (iii) data validation and cleaning tools to make data complete and accurate; (iv) metadata management tool to make data easy to find and understand. Figure 1 shows the architecture of our data platform. Next, we will use this data platform to validate and prepare the data.

Task 1.1: Prepare Infrastructure

This task consists in preparing the physical servers for deploying the data platform of the project. First, we install and configure the virtualization framework (i.e. hyper-V) on the servers. Second, we create Virtual Machines to deploy the data platform.

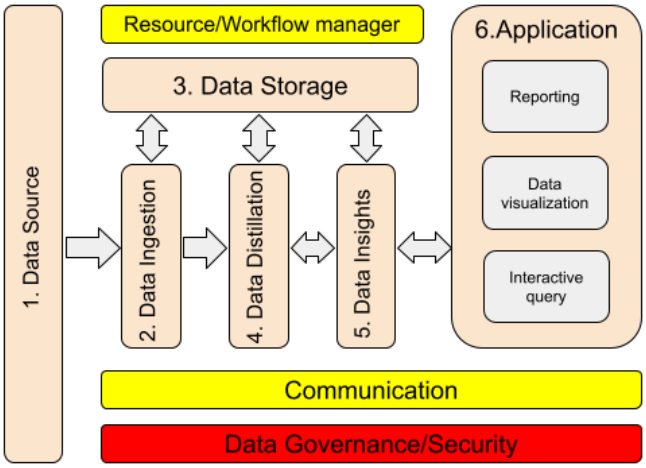

Figure 1. The architecture of the data platform

Task 1.2: Deploy data platform

This task consists in deploying the data platform. First, we deploy a distributed high-performance, s3-compatible object storage called Minio[1]. Second, we deploy a spark cluster[2] for executing data engineering, data science, and machine learning tasks. Third, we deploy and configure Jupiter lab to provide interactive notebooks to run tasks on the spark cluster.

Task 1.3: Data acceptance test

This task consists in validating data after we receive them from the CNAM. First, we develop an application that will validate the data. This application is structured around data validation rules. Second, we define these data validation rules with the domain experts. Third, we use the application and the included data validation rules to the received data. If one of the validation rules is not satisfied, we need to examine the origin of the error. If it is an extraction error, we need to contact CNAM to provide us with a new extraction.

Task 1.4: Data re-modelling

This task consists of transforming raw data into cleaned data. First, we address and fix errors detected during the acceptance test (task 1.3). Next, we sort the massive raw database by date of care instead of the date of reimbursement (date of flow) and create flag variables to identify rows to be deleted according to CNAM recommendation (e.g. code sex errors). To support the mass volume of the data, we will transform the existing SQL programs (provided by Constances) into pyspark scripts. Finally, we export the transformed data in parquet format[3], which is a column-oriented data file format designed for efficient data storage and retrieval.

Task 1.5: Infrastructure maintenance and help desk

This task consists in maintaining the accessibility of the data, the availability of the data platform, and providing assistance and information. This task will be carried out throughout the project.

Deliverables: Publication of the transformed data in format parquet.
Success indicators: Build an infrastructure able to perform "Data acceptance test" and "Data-remodeling" on SNDS extraction at nationalwide level.
Partners involved: CASD, UMS-011.

Work Package 2: Data Integration

Leader: S. KAB (UMS-011) and P. LIU (CASD)

The massive SNDS raw data sets present a challenge to researchers seeking to understand complex and varied information. The first objective of this work package is to create a single data table for each SNDS

[1] https://min.io/
[2] https://spark.apache.org/
[3] https://parquet.apache.org/

topic (eg. medication delivery table) by combining information from several raw tables considering good practices in linking and cleaning (CNAM recommendations). Furthermore, a sizable fraction of outcomes identification requires to combine several topics through algorithms (eg. identify lung cancer by combining specific drugs (ATC codes) and hospitalizations (ICD codes)). Thus, the second objective is to implement a set of algorithms based on the literature or expert recommendations (CNAM, ReDSiam) to be used in the data analysis.

Task 2.1: Create a simplified and advanced use SNDS

This task consists in using the pre-cleaned raw database obtained from Task 1.4 (WP1) already sorted by date of care. We will transform the existing SAS-SQL programs (provided by UMS-011 who already developed more than 20 topic tables [drug delivery, medical acts, medical consultations . . .]) into pyspark scripts.

Task 2.2: Identify specific diseases or indicators with SNDS algorithms

This task consists in using the topic tables obtained from task (2.1) to implement a set of algorithms based on literature or expert recommendations (CNAM, ReDSiam) and identify specific conditions based on SNDS healthcare consumption. We will transform the existing SAS-SQL programs (provided by UMS-011, more than 100 algorithms for 64 specific diseases/groups of treatments) into pyspark scripts.

Task 2.3: Data catalog

This task consists in building a detailed inventory of all data assets (e.g. raw data, transformed data, etc.) of this project. It is designed to help data professionals to quickly find the most appropriate data set for any given analytical purpose. First, we will develop scripts which can collect and parse existing metadata (e.g. data location, data owner, schema, etc.) of the data. Second, we store the collected metadata into a graph based data store, in order to build the index of the metadata, and data lineage. Third, we deploy a search engine with various filters based on the metadata and data lineage. All data professionals will be able to use this search engine to find the most appropriate data and understand them.

Task 2.4: Topic table and algorithms enhancing

During the analyses (see following WPs), when additional need is identified (e.g. a disease-specific cost table), a supplementary algorithm and the associated thematic table (or supplementary column) will be created.

Task 2.5: Methodological support and quality assessment

This task intends to: (1) provide methodological support to analytical teams, e.g. how to identify organized cancer screenings in the SNDS, (2) targeted expertise, e.g. impact of the new coding rules for dental procedures in 2014 (from NGAP [Nomenclature Générale des Actes Professionnels] coding to CCAM [Classification Commune des Actes Médicaux] coding).

Deliverables: SNDS topic tables (drug delivery ...), SNDS algorithms for specific pathologies and/or costs, Data catalog
Success indicators: methodological support to upcoming WPs, quality assessment of SNDS based outcomes
Partners involved: UMS-011, CASD.

Work Package 3: Machine Learning for Econometrics

Leader: G. LECUÉ (Essec)
Team members involved: A. Khaleghi, L. Wilner, A. Dalalyan, N. Chopin, Ph. Choné.

Task 3.1: Features engineering, outcomes construction and shocks identification

This task consists in providing a fine description of the healthcare network viewed as a series of bipartite graphs. These descriptions will be used during all our projects (see WP 3 and WP 4) either as **observed features** of the patients, healthcare professionals and the graphs, or as **outputs** that we may want to predict, or as **shocks**, such as the retirement of doctors that will be used to infer causal effects (see WP 4). We first focus on each side (patients then healthcare professionals) of the graphs, and then move to the very structure of the graphs themselves.

Patients features. First, patients will be characterized based on observed variables such as age, gender, geographic location, health conditions, healthcare utilization indicators, etc. We will pay particular attention to the geographic dispersion of health indicators and to the entry, move and exit of patients. Notice that individuals are present in the data only when they utilize healthcare resources, and are thus referred to as "patients". (We expect about 400,000 new patients every year.) For task 3.2, we need to know the structure of the population at the municipality level. For this purpose, we will resort to the census as well as to the data about local labor markets (e.g., about the structure of employment), amenities and equipment.

Healthcare professionals features. The data is unique in the sense that we observe all the patients of each provider and the entirety of their activity, which allows for a comprehensive view of their labor supply. Healthcare professionals will be characterized according to many dimensions, in particular their location, their specialty (generalist or specialist doctors, nurses, pharmacists, physiotherapists, etc.), their gender, and their activity. We will describe labour supply behavior, e.g., whether they work on Saturdays or at nights (night shifts are eligible to specific rates). Physician activity will be documented based on medical procedures and prescriptions (pharmaceuticals, sick leaves, etc.) as well as on the connections to other providers. Since the spatial aspect is crucial to our network analysis, we will need to properly visualize the geographic dispersion of professionals all over the French territory and describe how practice styles differ across geographic areas (see, e.g., Silhol 2020).

Graph features. Third, we wish to characterize the main features of the graph, which requires computing the degrees of all vertices, be they patients or professionals in the bipartite and the projected graphs. We will count the number of doctors visited by patients and the number of patients per year for each healthcare provider. We will closely examine patient-provider pairs, so-called dyads in Eliason et al. 2022's terms. Such an approach fully exploits the richness of the healthcare network data.

We will describe the distribution of the distances between patients and providers (healthcare professionals and hospitals), which will be measured for various metrics. Notice that unobserved dyads (patients and doctors who have *not* met) convey useful information to infer patients' travel costs. How to deal with the high number of unobserved dyads is part of the statistical challenges of our project; a starting point will be the recent methodological contribution in that domain by Gao et al. 2022, who resort to logical differencing as a way of canceling out the unobserved heterogeneity at the source of the network formation.

We will measure how much the care of a given patient i is dispersed across many providers. For patient i, Agha, Frandsen, et al. 2019 define the index of patient care continuity as

$$\text{PCC}_i = \sum_p \left(\frac{n_{ip}}{N_i}\right)^2$$

where n_{ip} is the number of visits to physician p and $N_i = \sum_p n_{ip}$ is the total number of visits.

We will also carefully characterize the existing links between various providers through their patients – the so-called "projected graphs". Doing so will allow us to describe the patient-sharing networks. Following the recent literature on referral decisions, e.g., Agha, Ericson, et al. 2022, Zeltzer 2020, we can quantify, from the point of view of some primary care provider (PCP) d, the concentration of specialists based on the Team Referral Concentration (TRC) index:

$$\text{TRC}_{dj} = \sum_s \left(\frac{m_{ds}}{M_{dj}}\right)^2$$

where m_{ds} accounts for the number of patients of d shared with specialist physician s while $M_{dj} = \sum_s m_{ds}$ is the number of patients of d shared with any specialist.[4]

In all of the above analyses, it will be necessary to carefully distinguish between attending the default physician ("médecin traitant") and physicians who may be consulted on a particular instance.

[4] By specialist, we also mean nurses, pharmacists, etc.

Outcomes. Fourth, we will compute health outcomes. A non-exhaustive list of corresponding indicators that can be found in the literature includes healthcare utilization, namely annual spending per patient, or the cost covered by National health insurance, Currie et al. 2020; Dafny 2005, healthcare providers' activity, length of stay for hospital admissions, (potentially inappropriate) use of emergency departments, see Agha, Frandsen, et al. 2019; Bruni et al. 2016; Finkelstein et al. 2016; Gaynor et al. 2013; Gottlieb et al. 2010; Shen 2003; Tay 2003)

An important empirical challenge consists in assessing the quality of care, which is highly multidimensional. We will construct process measures such as appropriate management of chronic diseases, imaging, testing, preventive care. We will identify adverse health outcomes based in particular on ICD codes in PMSI (see, e.g., Avdic et al. 2018). and (potentially avoidable) hospital readmissions. Another classic avenue of research is to proxy hospital quality with mortality rates (especially following acute myocardial infection), see Cooper et al. 2011; Currie et al. 2020; Dafny 2005; Finkelstein et al. 2021; Gaynor et al. 2013; Kessler et al. 2000; Propper et al. 2010.

Shocks. Fifth, it will be possible to characterize how the healthcare network varies over time, hence to focus on how previous distributions of, say, degrees and health indicators change and co-vary over the period. For instance, annual distributions of degrees provide the researcher with a measure of how the structure of the graph changes over time. In particular, and to anticipate WP 4 below, we will seek to isolate shocks that affect the structure of the healthcare network. Of crucial interest in this regard is any entry or exit of both patients (e.g., childbirths, deaths, or when individuals move) or doctors (e.g. any relocation, on top of death and retirement). Of course, the same reasoning applies to other healthcare providers like hospitals, with closures of some units (e.g., maternity wards) providing us with an exogenous source of identifying variation for the causal effect of the graph on health, see Chatterji et al. 2023, Avdic et al. 2018, Avdic et al. 2019 and Bergonzoni et al. 2021.

Data engineers from UMS-011 will bring their expertise on the data (for instance, they will identify screening tests to document preventive care, help us in the computation of costs incurred by National health insurance, etc.).

Task 3.2: Accessibility indicators and medical deserts

The geography literature has developed accessibility indicators, in particular the two-steps and three-steps floating catchment areas (FCA) indicators, 2SFCA and 3SFCA in short, see Lucas-Gabrielli et al. 2022; Jun Luo 2014; W. Luo et al. 2009; PL. 2013; Wan et al. 2012. We will enrich them by estimating in a flexible manner the probability $\mathbb{P}_{ij}$ that a patient located in area i visits a provider located in area j, see Figure 2.

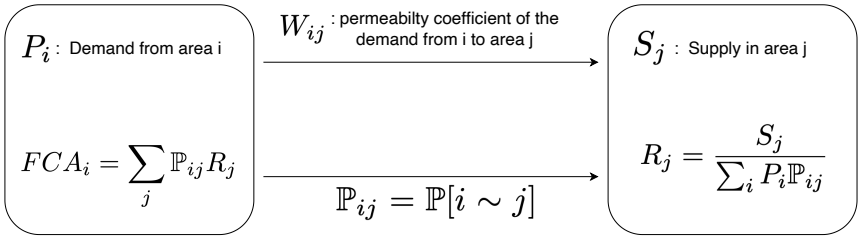

Figure 2. P_i and S_j represent the populations of potential patients in area i and of providers in area j. The coefficients W_{ij} decrease with the distance between i and j. $\mathbb{P}_{ij}$ is the probability that a patient from i visits a provider in j. R_j is the supply per patient in j. FCA_i is the medical density in i.

FCA methods differ in the specification of the choice probability $\mathbb{P}_{ij}$: for 2SFCA, $\mathbb{P}_{ij}$ depends only on the distance between i and j through the specified function W_{ij}; for 3SFCA, $\mathbb{P}_{ij}$ depends on the distance-weighted number of providers in j, $W_{ij}S_j$. The probability may instead depend on the distance-weighted number of providers *per patient* in j, $W_{ij}R_j$, because what matters for patients is not the gross number of providers in a certain area but the availability of these providers which is reflected in R_j. As R_j itself depends on $\mathbb{P}_{ij}$, computing the medical density indicators requires solving a fixed point problem.

Our main goal is to learn the probabilities $\mathbb{P}_{ij}$ from the data. To that end, we will rely on the vast literature on graph modeling and link formation H. Wu et al. 2022 to model this probability of connection and learn it using a machine learning approach. We will also use features influence tools (such as the Shapley values Rozemberczki et al. 2022) to quantify in a data-driven way, the role of the distance on the probability of connection between patients and doctors without having to guess or set it a-priori.

In standard FCA models, the probabilities $\mathbb{P}_{ij}$ sum up to one. Yet an important issue concerns those individuals in the population P_i who do not utilize the considered type of health care. Discrete choice models generally include an outside option that represents all the options non modelled. (For instance, when studying the choice of a GP, one might include other professionals, or hospital care, or no care at all, in the outside option.) It is possible to infer from the data the share of the population P_i who is concerned with the issue, Choné and Wilner 2022; Dubois et al. 2018. This allows for the identification of the set of *potential patients*, as opposed to the whole population within some location.

We will also implement a label propagation algorithm (a semi-supervised clustering method) under geographic constraints to infer medical deserts (see L. Bai et al. 2020). We may first find initial labels (i.e. 'medical deserts' and 'good supply area') from one of the FCA models and keep only the areas with medical density supply below the $1/100$ quantile and the one above the $99/100$ quantile. We will then spread the labels all over the territory using a label propagation algorithm taking into account the distances between areas as a geographical constraint.

Task 3.3: Patients and healthcare providers embeddings

Task 3.2 relies only on observed patients' and providers' characteristics. Yet healthcare utilization is the outcome of multiple decisions where unobserved characteristics certainly play a major role. On the patients' side, health status and medical needs are 'hidden' features that reveal themselves indirectly in the data. Medical needs are only one driver of the demand for care. The patients' propensity to consume medical services, to ask for multiple medical opinions, to favor more intensive care, etc., depends on individual preferences (e.g., attitude to risk), habits, and beliefs. The same is true on the providers' side. We do not observe the providers' cost and financial incentives, their marginal utility of income and wealth, their household composition, their opportunity cost of time, their reputation, etc. Physician decisions regarding treatment choices also depend on their beliefs and preferences. Cutler et al. 2019 refer to doctors who tend to opt for more (less) aggressive care as 'cowboys' and 'comforters' respectively.

In this task, we aim to infer unobserved patients' and doctors' characteristics from the data, using the shape of the patient-provider links, the projected graphs (patient-sharing network) and from prescriptions and medical procedures. The econometric and ML approaches refer respectively to the notions of **fixed effects** and **latent variables**. The strategies that will be developed in this work package apply to many setups depending on the choice of features and outcomes. We present below a glimpse of them only when the considered outcome is the existence of a link between a patient and a professional.

Fixed effects in the logistic model for adjacency matrix modelling. In the case where the outcome $\mathbb{Y}_{ij}$ is the existence or not of a link between a patient i and a doctor j - with observed features $\mathbb{X}_{ij}$ - a natural model is a logistic model with fixed effects, see Abowd et al. 1999; Bonhomme 2020:

$$\ln\left(\frac{\mathbb{P}[\mathbb{Y}_{ij}=1|\mathbb{X}_{ij}]}{\mathbb{P}[\mathbb{Y}_{ij}=0|\mathbb{X}_{ij}]}\right) = \langle \mathbb{X}_{ij}, \boldsymbol{\beta}^* \rangle + \alpha_i + \psi_j \tag{1}$$

where α_i and ψ_j are the (unobserved) fixed effects that can be interpreted as quantitative measures of patient i utilization of health services and health provider j willingness to provide healthcare offer. In particular, the α_i's are the (quantitative) parameters that we want to estimate regarding our objective.

We will estimate $(\alpha_i)_i$ and $(\psi_j)_j$ in the SNDS database. We may also slice our database according to some geographical data (ex.: all patients living in some given area) or according to doctors' specialities or economic parameters.

The machine learning viewpoint on hidden features is that of latent variables. We will use machine learning techniques to quantify unobserved demand for healthcare with latent variables in order to overcome the limitations of the standard AKM model (1): the dependency between the feature vector $\mathbb{X}_{ij}$ and the

output is linear; the effect of hidden features of i and j are measured by a scalar (a real number) whereas we explained above that the unobserved heterogeneity in patient and provider preferences is multidimensional; there are no interactions between the features (hidden and/or observed) of i and j. In particular if i has the tendency to visit mostly female doctors then such a feature of the couple (i, j) would not be taken into account in model (**1**).

Based on a machine learning approach, we introduce more complex models that will explain in a better way the graph formation by taking into account more complex non-linear, direct and cross effects of hidden and observed features. An important outcomes of these models are the construction of embedding layers that will give us access to quantitative measures of patients demands and healthcare providers offers: the latent variables that we are looking for. Indeed, we believe that a single number can hardly summarize in its own such multi factorial concepts as the one considered here. We are therefore looking for a '**vectoriation**' of patients regarding some outcome(s).

Embedding layers of patients and health care providers. When the output $\mathbb{Y}_{ij}$ is the existence or not of a link between i and j, the prediction problem is a binary classification problem which aims at explaining the graph formation. This graph formation is explained in (**1**) by the (observed) features vectors $\mathbb{X}_{ij}$s and the fixed effects $(\alpha_i)_i$ and $(\psi_j)_j$. A first extension of this model that we will study is to consider latent vectors as in recommendation systems (see Mnih et al. 2008; Salakhutdinov et al. 2007; Shan et al. 2010). We will then extend this approach to construct non-linear models with cross interaction between observed and hidden features. The general model for the graph formation problem is to predict the probability of output $\mathbb{Y}_{ij}$ to be 1 (i.e. existence of a link) by a complex (non-linear) function of the observed features $\mathbb{X}_{ij}$ and some latent vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\psi}_j$. We will use artificial neural network (ANN) to construct this type of model. Our first aim is to learn the latent vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\psi}_j$ (for all patients i and healthcare providers j) and the key step for that is the patients and health providers embedding layers as depicted in Figure 3.

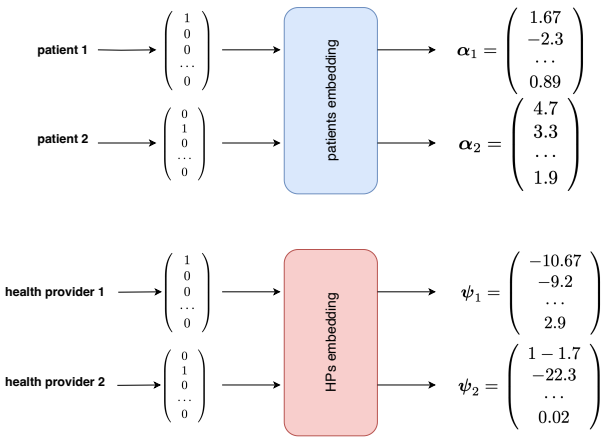

Figure 3. Vector representation (i.e. vectorization) of patients: '*patients embedding*' and health providers: '*HPs embedding*'. Both embeddings are (linear or non-linear) applications which take as input one-hot encoding vectors (i.e. vectors of the canonical basis) of $\mathbb{R}^N$ ($\mathbb{R}^J$ resp.) where N (J resp.) is the number of patients (health providers resp.) and goes into $\mathbb{R}^\ell$ where ℓ is the dimension of the latent space to choose - we may choose different dimensions for the two embeddings. Embedding layers are learned during the learning step of the entire ANN (see Figure 4 below). Once learned, the two embedding layers will provide the two families of vectors $(\boldsymbol{\alpha}_i)_{i\in[N]}$ and $(\boldsymbol{\psi}_j)_{j\in[J]}$ that, together with the observed features, will be used during all the project (see WP 4).

The two embeddings layers (one for the patients and one for the healthcare providers (HPs)) are transformation of a patient (HP resp.) represented by a one-hot encoding vectors into a vector of a certain size as shown in Figure 3. We will do two such embeddings in the first step / layer of the neural network we are going to build. Dimensions of embeddings usually represent the space needed to represent complex concepts associated to our prediction task such as here 'the demand for health' (on the patients side) and the 'health offer' (on the HPs side). A general overview of the artificial neural network (ANN) structure we will use is represented in Figure 4, see Mikolov et al. 2013; Torregrossa et al. 2021.

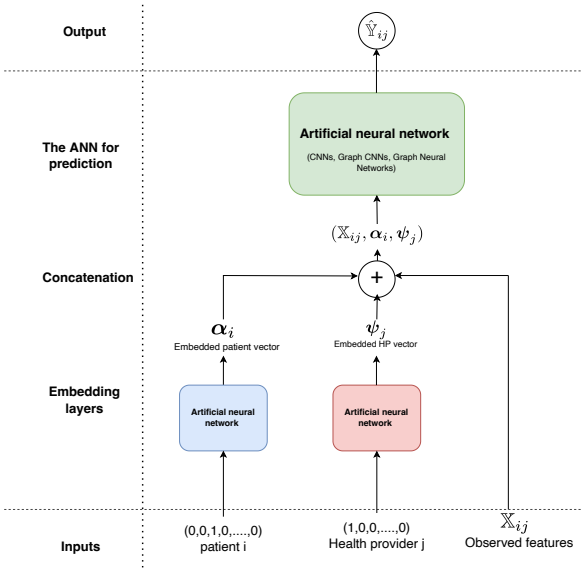

Figure 4. The general structure of artificial neural networks for the construction of a machine learning model that predicts an output $\hat{\mathbb{Y}}_{ij}$ from the one-hot encodings vectors of patient i and HP j and their observed features $\mathbb{X}_{ij}$. Our primary goal is to retrieve the two embedding layers from this network.

The heart of this work package will be the construction of the two embedding layers, one for patients and one for healthcare providers. We will store these two layers as tools that will be used for latter purposes such as clustering patients and assessing medical demand and medical offer. The construction of these tools is key to our project. We will therefore construct several ANN architectures (see S. Wu et al. 2022) and assess their quality using various KPI (key performance indicators), some of them coming from recommendation systems (see Bobadilla et al. 2013). Given the very large size of our bipartite graph at the national scale, we will have to use some framework devoted to this massive amount of data (see for instance NVIDIA Merlin framework and its Sparse Operation Kit) to train deep learning network on this database.

We will also explore a third path lying in between the econometric AKM model and the agnostic learning setup of neural networks. This path goes through the concept of hidden variables used in the Bayesian literature (see Chopin et al. 2020). Fixed effects may be looked as random variables (see Bonhomme 2020) whose distributions will be estimated via variationnal inference (see Jordan et al. 1999) and message passing algorithm on the bipartite graph (see Bayati et al. 2013).

Task 3.4: Time series models for the statistical & causal analysis of the network

The main objective of this task is to design and use advanced time series models which take the spatio-temporal structure of the data into account. Particular attention will be paid to interpretability of the results. Using sliding windows and methods based on block-bootstrapping, and by constructing appropriate LSTM architectures (see Hochreiter et al. 1997), we will extend the methods developed as part of Tasks 3.1-3.3 to generate time series of graph-based features of the network. The time series so-developed will naturally have long-range dependencies. We will adopt provably effective time-series clustering and change-point estimation algorithms such as those of Khaleghi and Ryabko 2020; Khaleghi, Ryabko, et al. 2016, and Khaleghi and Ryabko 2012, 2014, 2016 (see also Brodsky 2016 and references therein), which come with statistical guarantees for dependent sequential data. Given the objective to detect and understand 'medical deserts', we will further extend the algorithms to simultaneously take the spatial aspect of the data into account. Moreover, inline with the vision of WP 4 we will build upon the algorithms of Khaleghi and Lugosi 2023 to develop statistical tests for the predictive causality (in the sense of Granger 1969) of the impact of the healthcare network on health outcomes. To achieve this objective, we will rely on the technical results of Grünewälder et al. 2021 to consistently estimate function-valued conditional expectations. Finally, inspired by policy-evaluation techniques of the Reinforcement Learning community (see, e.g. Meyn 2022), we will adopt some of the (restless) Multi-Armed Bandit methods of Grünewälder et al. 2019 to develop algorithms for the retrospective evaluation of the efficacy of the existing healthcare system policies.

Remark: We will first assess the performance of our algorithms on simulated data produced by toy models of the SNDS before turning to the entire 'real-world' database. This way, we will not be bound to wait for the construction of the platform delivered by our team members from CASD to test and calibrate our statistical methods.

Deliverables: Publications in top-tier machine Learning, statistics, econometrics and medical journals and conferences.
Success indicators: scientific publications.
Partners involved: CREST, CASD.

Work Package 4: Causal Effects of Healthcare Networks and Medical Deserts
Leader: F. KRAMARZ

Team members involved: Ph. Choné, P. Leveneur, B. Schmutz, A. Uhlendorff, L. Wilner, G. Lecué and A. Khaleghi. We also intend to collaborate on this work package with colleagues who are expert in health data: Pierre Dubois (Toulouse School of Economics), Sylvie Blasco and François Langot (Le Mans Université), and Johan Vikström (University of Uppsala).

Our primary objective is to infer the causal impact of the structure of the French healthcare network (HN, hereafter) on health outcomes. This objective directly is a scientific reformulation of the prominent issue of 'medical deserts' and their consequences on health of citizens within the French public debate.

To achieve our scientific goal, a key component of this endeavor is to isolate shocks that will affect both the HN and the health outcomes only through their effect on the HN. Examples of such sources of identifying variation comprise shocks that are arguably exogenous to patient health, but that do modify the structure of the network by creating, removing, or displacing vertices of this HN, altering among other things the structure and distribution of its degrees over time. For instance, patients may die, or move across locations. Physicians themselves may also die, retire from the labor market, or move across locations. Again, these events should not be caused by the agent's health status (except for death, admittedly). Other types of shocks often used in the literature come from "displacement" events. In our case, the data will allow us to observe entry and exit of various healthcare providers, such as hospital openings and closures (see Task 3.1 and Avdic et al. 2018, Avdic et al. 2019).

The data accessible to our group will allow us to systematically track any change in the structure of the HN (see Task 3.1). It will for instance enable us to follow patients over time, possibly from a region to another, based on his individual identifier (an anonymized version of the NIR), which is both patient-specific and time-invariant, but also thanks to the location of his healthcare providers. Similarly, it will be possible to infer from the data when a physician dies or retires from the labor market, based on a physician-specific identifier. We will then be able to observe how the set of her former patients is referred to other physicians at some point in time, and how much time this referral takes. Similar event-based approaches can be used to identify hospital closures, etc.

After a vertex has disappeared from the HN – some general practitioner (GP) retiring from the labor market, say – we will observe the reallocation of patients across the set of (remaining) GPs. We will thus seek to assess on which basis such link formation occurs, namely depending on which observed characteristics, including patient-specific, physician-specific, patient-physician (dyad) observed characteristics like distance, or any relevant feature of the HN such as those used in the literature (see, e.g., gender homophily of links between doctors and hospitals, Linde 2019, Zeltzer 2020 on this topic). A particular challenge will be to find statistical strategies allowing us to go beyond such observed characteristics to include unobserved ones, especially assuming that these unobserved components are time-invariant. Clearly, machine-learning techniques will be of utmost importance when trying to identify the (vectors of) bundles of characteristics that matter, see WP 3.

To give a sense of how these techniques could be used in our context, think about finding a set of observed variables (gender, age, location as well as various links to other MDs or health professionals in the past allocation, i.e. the structure of the pre-event HN, see features engineering Task 3.1) that could help predict the post-event allocation. Related questions are the role of links with a longer, or shorter, duration, but also groups of smaller sizes shaping the post-event allocation, etc. Given our unusual measurement of such links, a particular emphasis could be placed on the role played by various healthcare players such as nurses or pharmacists in the reallocation process. Do we observe any heterogeneity with respect to patient

characteristics in the resulting matching process? How informative is the time elapsed since a physician exited the market, and before which patients are assigned to a new doctor? This part of our research places particular emphasis on patient-physician pairs viewed as dyads, which very few other analysts can access in contrast to the data variables we can leverage, and following what has been done by Eliason et al. 2022 between workers and firms. Importantly, this approach relies on unobserved factors at the dyad level, which builds upon previous methods considering only patient- and physician- individual fixed effects as in Abowd et al. 1999, hereafter AKM, or in Graham 2017 where those latent components are assumed to interact additively –an assumption that may be relaxed based on machine learning techniques, though (see WP 3).

At this stage, we have not yet discussed causality of the HN on individual health. Indeed, using a Rubin-style causal setting, we intend to place ourselves within a model in which health outcomes causally depend on connections within the HN, viewed as the treatment. Because the network structure is endogenous in that it depends on both supply and demand of healthcare, an identification strategy we will consider is instrumentation using the exogenous shocks mentioned above, in the tradition of the IV literature. Another possible identification strategy will be based on a difference-in-difference (DinD) approach, inspiring ourselves from the recent literature on staggered adoption designs (De Chaisemartin et al. 2020, dCdH hereafter). The treated group will comprise all patients faced with, say, retirement of their GP at a given date t. In dCdH, the control group for a treated group at date t will comprise all patients exposed at a later date to the same shock (e.g. retirement of their own GP). To make things fully comparable, we will condition on various effects, again leveraging techniques inspired from machine learning.

Task 4.1: Construction of control groups with clustering techniques

Clustering techniques from Machine learning will help us design control groups in order to evaluate the causal impact of HN over an outcome (or a group of outcomes) based on shocks. Assume for now that we want to exploit the retirement of some GP as a shock. We first compute an embedding of all patients related to the health outcome(s) considered (see Task 3.3). This provides an embedding vector for each patient that quantifies the hidden features of each patient related to the outcome(s) in a multivariate way. Then, for each patient, the patient embedding is concatenated with his/her own vector of observed features (the one that has been constructed after Task 3.1). Each patient is now associated with a vector made of hidden features (learned via the network embedding) and of observed ones. Next, we cluster all these vectors (the number of vector is the size of the French population if the largest scale is considered). The econometric interpretation of those clusters is related to the Common Trend Assumption (CTA) since all patients within a given cluster have a common hidden and observed pre-trend. Following dCdH, we may next split each cluster into treated sub-groups depending on the year of retirement of their GP, plus some control group composed of patients whose GP did not retire (see Figure 5 below, which shows this partition of patients). This approach can be extended in many ways such as other clustering than the one of patients (for instance dyads or triads), other shocks, other attributes than living in a medical desert or not (e.g. attributes on the projected graph of doctors, and related to their organization, etc.), etc.

In addition, these shocks should help us identify unobserved health needs, especially in the case when the observed utilization is constrained, due to true demand being rationed by, say, the lack of healthcare supply in a 'medical desert'. Again, we intend to use machine learning techniques to help us recover counterfactual utilization, i.e. that which would have prevailed if demand were not constrained by limited supply. Such measures will be confronted to alternative measures that directly stem from administrative categories ("Zones d'Intervention Prioritaire", ZIP hereafter) and allow us to contrast our definition of a medical desert with the administrative view (see https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/lindicateur-daccessibilite-potentielle-localisee-apl for a description).

Task 4.2: Construction of predictive models to assess the effect of medical deserts

To illustrate, let us consider some sub-division of the population into three groups depending on healthcare supply: medical desert, medium medical supply, good medical supply (See Task 3.2 for a quantitative measure of medical density). We focus only on the two extremes: 'medical desert' and 'good medical supply'. We now consider some outcome(s) and the related embedding vectors of every patient related to this outcome(s) (see Task 3.3 for such patients embedding). Remember that we interpret the

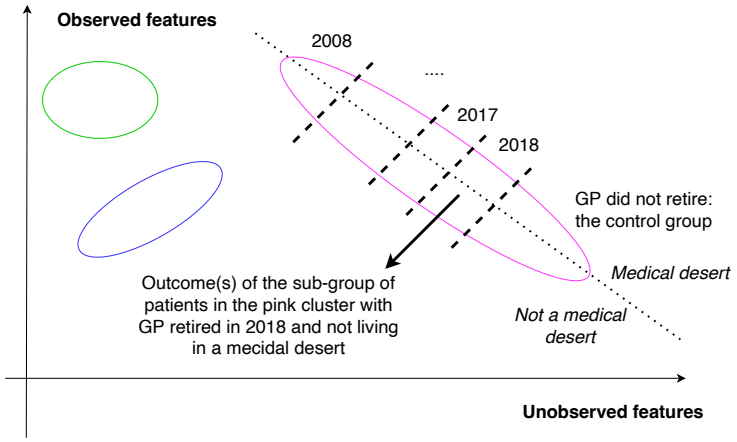

Figure 5. Patients have been clustered in the cross space of hidden and observed features (for each patient, hidden features have been learned via an embedding layer for the same output(s) as the one under study, see Task 3.3). Each cluster is split according to the year of the shock (from 2008 to 2018) and a sub-group of patients who did not face such a shock. Here the shock is the retirement of the GP. Finally, within each sub- group we identify those who live in a medical desert or not. Outcomes are then collected for each (sub-)sub-group and compared with the ones of the controlled group.

embedding vectors as multi factorial quantities representing hidden features of patients related to the considered outcome(s). We now train two predictive models of the outcome(s): one for each of the two sub-populations, living in a medical desert or not. Next, we identify individuals moving from one group to another, and we compute the difference in corresponding outcomes, i.e. the one actually observed and the counterfactual one (the latter being predicted using one of our two predictive models, depending on the type of area where he/she was living before). That difference should be a measure of the causal effect of the living area being a medical desert on the outcome. We have here described the general idea behind our causal estimation; the implementation may follow some IV or some generalized DinD strategy in a dCdH spirit as already mentioned.

In contrast with methods used in labor market or in international trade mentioned previously, in which wages or prices tend to be the primary outcome of interest in the latter examples, our setting requires to consider multiple outcomes (see Task 3.1) possibly explained by multiple dimensions of observed or unobserved covariates (see Task 3.3). The specifications used in these settings may lack flexibility and are likely to rely too heavily on linear models (including linear probability models, see Task 3.3). Hence, we will likely need to adopt a multivariate framework with latent vectors to capture unobserved components of the processes under consideration as well as non-linearities as in Task 3.3. Such observed or unobserved (latent) vectors/features may refer to physician practice styles, beliefs or preferences (from either side, i.e. both patients and doctors).

Deliverables: some reports.
Success indicators: scientific publications.
Partners involved: ESSEC, CASD, CREST, UMS-011.

II Organisation and implementation of the project

a Scientific coordinator and its consortium / its team

GRAPH4HEALTH gathers a unique mix of expertise in data science (statistics, machine learning, and econometrics), economics, epidemiology, and big data analytics. The consortium puts together:

- Center for Research in Economics and Statistics (CREST, CNRS UMR 9194), specifically its statistics team and its economics team;
- Centre d'Accès Sécurisé aux Données (CASD);
- UMS-011, Unité Cohortes épidémiologiques en population, INSERM ;

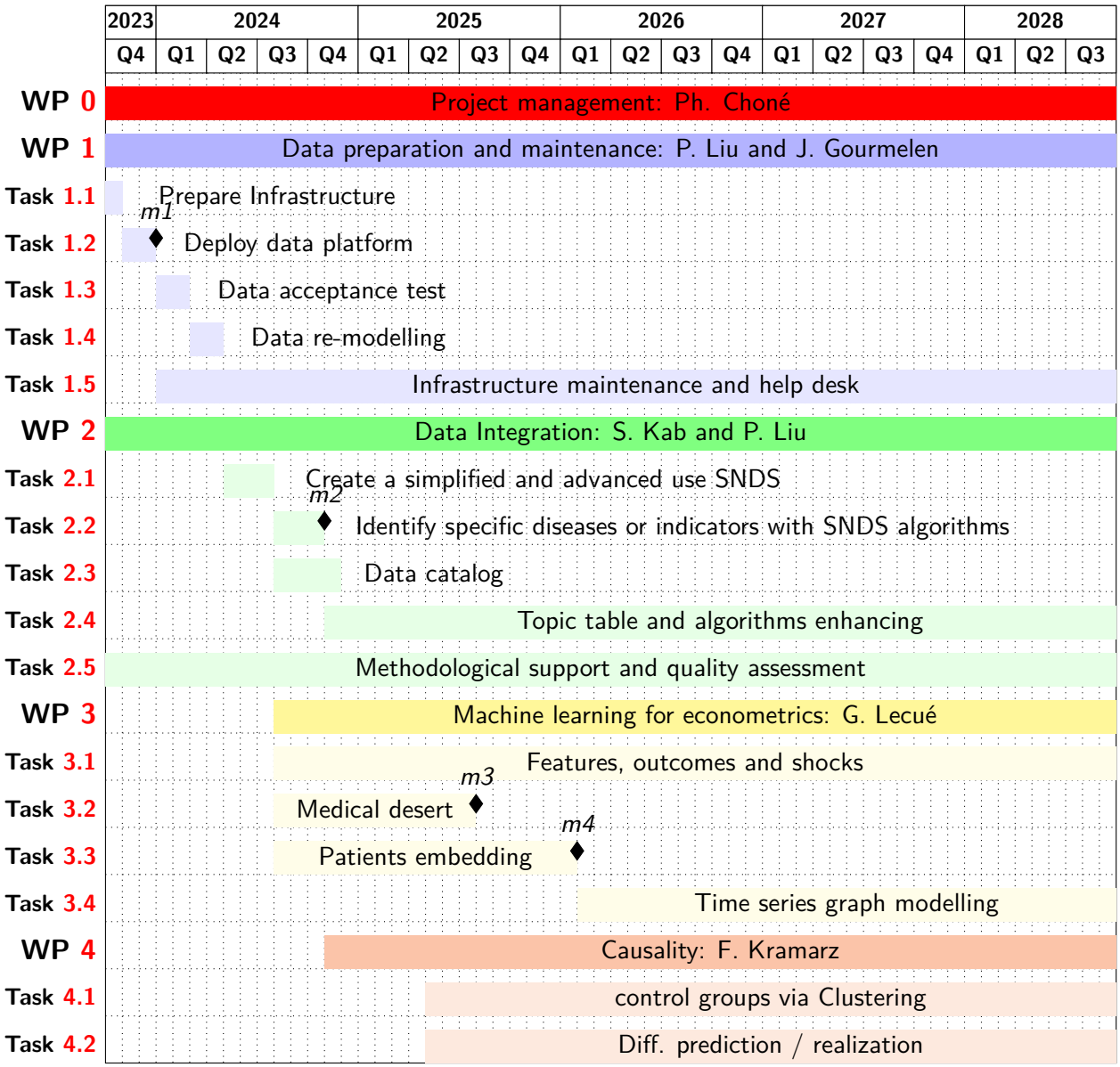

Figure 6. Gantt Diagram of the project. We have identified four milestones in our project: **m1** is when the platform will be made available to us, **m2** is when simplified tables and SNDS algorithms will be released, **m3** and **m4** are when medical densities and a patients embedding will be constructed. They are the prerequisites to assess the causal effect of the graph network organization (in particular of medical deserts) on outcomes. Tasks 1.3, 1.4, 2.1 and 2.2 will be repeated after each new delivery of data. We expect two more deliveries (after the one of March 2023) during the five years of the project.

- ESSEC business school.

Graph4Health is coordinated by Philippe CHONÉ, Professor of economics at CREST and Head of Research of Groupe des Ecoles Nationales d'Economie et de Statistique. He holds a PhD in mathematics and a Habilitation à Diriger des Recherches in economics. He has published articles about physician behaviour and hospital regulation Choné and Wilner 2022, Choné, Coudin, et al. 2020, Choné and Ma 2011, some of which are based on the same data as the present project. Together with a PhD student, he is already collaborating with the UMS-011 team on a related project that uses the Constances epidemiological cohort.

The project is multidisciplinary at its core. CREST is a leading center in economics and statistics. Philippe Choné, Francis Kramarz, Lionel Wilner, Pauline Leveneur, Arne Uhlendorff, Benoît Schmutz bring expertise in economics and applied econometrics.

Guillaume Lecué is a professor of Statistics and Machine Learning at ESSEC. He has published articles on matrix completion and community detection (Chrétien et al., 2021). He has worked on several data

Researcher	Person.month	Call, funding agency, grant allocated	Project's title	Scientific coordinator	Start–End
Ph. Choné	6	ERC	Firms and Their Networks	F. Kramarz	2019-2024
G. Lecué	7	ANR	ADDS	Nabil Mustafa	2019-2023

Table 1. Implication of the scientific coordinator and partner's scientific leader in on-going project(s)

science projects on large-scale databases for private companies. Nicolas Chopin (Crest), Arnak Dalayan, and Azadeh Khaleghi are all experts in statistics and machine learning.

UMS-011 has over 10 years of SNDS data expertise. Sofiane Kab is pharmacist and holds a PhD in epidemiology and public health. He is co-head of two work packages, "Data innovation" and "methods and scientific advice" at the Constances cohort.

CASD has a strong expertise in big data analytics and metadata management. CASD data engineers are used to manage health data as well as hugely voluminous data sources in fields such as education, justice, taxation. A central issue when treating healthcare data is data protection and government regulation compliance. As CASD is a certified healthcare data hosting center ("hébergeur de données de santé"), we will have all the necessary support for data protection and regulatory compliance.

Two partners of Graph4Health have already developed an application that enables SNDS extraction to be validated (checking the presence of all the requested tables/variables, years of the flows, formats, etc.). In addition, ad hoc programs structure the raw SNDS tables by dates of care (from flow dates) and create a relational model. The new platform will enable us to represent the data as a series of time-evolving, geolocated, and bipartite graphs, with metadata on nodes and links (e.g., patients' health and doctors' characteristics). The graph perspective on the data will bring unique insights on the functioning of the healthcare system. Modelling the network formation and the causal effects of the graphs will allow us to better understand both the effect of care accessibility on health outcomes and the response of the healthcare system to the population's medical needs.

.

III Impact and benefits of the project

Informing the public debate: According to the Caisse nationale de l'Assurance Maladie, about 6 million French citizens, among whom 10% have chronic diseases, are not assigned an attending physician ("*médecin traitant*".) Poor accessibility may lead patients to give up on healthcare. The risk of deserting healthcare has prompted a lively policy debate in France on the ways and means of improving care accessibility (by better organizing the primary care sector, HCAAM 2022, by choosing between financial incentives and mandatory provisions for providers' location choices, Duchaine et al. 2022, etc.). Graph4Health will inform the debate by better characterizing the notion of medical deserts and by quantifying their causal effects on health care utilization.

Transferring technologies: Graph4Health will contribute to the dissemination of machine learning techniques and big data analytics among healthcare researchers. We will develop collaborations with the health services research community, in particular with IRDES in France. We also intend to work in close contact with the Direction de la statégie, des études et des statistiques of Caisse nationale de l'Assurance Maladie, which produces and studies the data.

Producing innovations: The collaboration between economists and statisticians should yield new results in the nascent field of causal machine learning. The main synergy between ML and econometrics in our context will be to evaluate the causal impact of the network dynamics on health care utilization (using control groups constructed from a clustering algorithm based on observed and unobserved features). New machine learning techniques will be developed to study healthcare networks seen as large-scale bipartite graphs. This will allow to finely describe the time-evolving networks and to develop new clustering techniques adapted to bipartite graphs. We also expect scientific spillovers in mathematical statistics, e.g., theoretical results for various estimators of fixed effects in AKM models based on adapted logistic LASSO or sparsity

methods under various graph modelling assumptions. We will disseminate our results in renowned machine learning conferences such as ICML and NeurIPS which have sessions and workshops on machine learning and health.

Publishing in scientific journals: To get a better sense of our publication and dissemination goals as a team, we clearly wish to publish papers in Journals of the highest quality, within the fields of specialization of all the team members, statistics, economics, epidemiology, and more generally in general interest scientific journals as well as medical journals. The topics we cover are regularly treated in such medical journals (e.g., Barber et al. 2017; Mahase 2019; Mayor 2018; Ni et al. 2022), maybe with less emphasis on the type of data and methodological approaches than adopted here.

Contributing to open sources: Finally, Graph4Health will make open-access code and results available to the community of researchers, practitioners, and policy makers interested in the health care system. This project is the first of this scale since SNDS was made available for research projects. (1) The architecture of the health data platform and its implementation will be reused for other following healthcare data research projects. (2) This project will also demonstrate how to store, manage, and analyze data of big volume efficiently (spark architecture and code). (3) A complete and clear procedure and programs for all necessary steps to deal with SNDS massive database that range from: acceptance test, data-remodeling by date of care instead of the date of reimbursement, creation of topic tables (drug delivery, medical acts, medical consultations, hospitalizations, long-term illness, etc.) and specific diseases identification (cancer, cardiovascular, psychiatric conditions, etc.) and/or indicators (total costs, specific costs, screening tests, etc.). This will cover most of other research project needs, it represents more than 10 years SNDS experience accumulated by Constances' teams and all improvement that will be made through this multidisciplinary project (e.g., new indicators for economic studies) available to the SNDS community. (4) The project will build the data catalog of the SNDS raw and transformed data which can be reused by other researchers and projects. The data catalog will provide a full-text search engine to quickly find the most appropriate data for any research purpose.

IV References related to the project

References

Abowd, John M, Francis Kramarz, and David N Margolis (1999). "High wage workers and high wage firms". In: *Econometrica* 67.2, pp. 251–333.

Agha, Leila, Keith Marzilli Ericson, Kimberley H Geissler, and James B Rebitzer (2022). "Team relationships and performance: Evidence from healthcare referral networks". In: *Management science* 68.5, pp. 3735–3754.

Agha, Leila, Brigham Frandsen, and James B Rebitzer (2019). "Fragmented division of labor and healthcare costs: Evidence from moves across regions". In: *Journal of Public Economics* 169, pp. 144–159.

Avdic, Daniel, Petter Lundborg, and Johan Vikström (2018). "Mergers and birth outcomes: evidence from maternity ward closures". In.

— (2019). "Estimating returns to hospital volume: Evidence from advanced cancer surgery". In: *Journal of health economics* 63, pp. 81–99.

Badinski, Ivan, Amy Finkelstein, Matthew Gentzkow, Peter Hull, and Heidi Williams (2022). "Geographic Variation in Healthcare Utilization: The Role of Physicians". PhD thesis, Chapter 3, MIT.

Bai, Liang, Junbin Wang, Jiye Liang, and Hangyuan Du (2020). "New label propagation algorithm with pairwise constraints". In: *Pattern Recognition* 106, p. 107411.

Baldominos, Alejandro, Fernando Rada, and Yago Sáez (Mar. 2017). "DataCare: Big Data Analytics Solution for Intelligent Healthcare Management". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 4, pp. 13–20. DOI: 10.9781/ijimai.2017.03.002.

Barber, Ryan M, Nancy Fullman, Reed JD Sorensen, Thomas Bollyky, Martin McKee, Ellen Nolte, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M Abbas, et al. (2017). "Healthcare Access and Quality Index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015". In: *The lancet* 390.10091, pp. 231–266.

Bayati, Mohsen, David F Gleich, Amin Saberi, and Ying Wang (2013). "Message-passing algorithms for sparse network alignment". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 7.1, pp. 1–31.

Bergonzoni, Alice and Marion Simon (2021). "La part des femmes en âge de procréer résidant à plus de 45 minutes d'une maternité augmente entre 2000 et 2017". In: *DREES* Études et Résultats 1201.

Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez (2013). "Recommender systems survey". In: *Knowledge-based systems* 46, pp. 109–132.

Bonhomme, Stéphane (2020). "Econometric analysis of bipartite networks". In: *The econometric analysis of network data*. Elsevier, pp. 83–121.

Brodsky, Boris (2016). *Change-point analysis in nonstationary stochastic models*. CRC Press.

Bruni, Matteo Lippi, Irene Mammi, and Cristina Ugolini (2016). "Does the extension of primary care practice opening hours reduce the use of emergency services?" In: *Journal of Health Economics* 50, pp. 144–155.

Chatterji, Pinka, Chun-Yu Ho, and Xue Wu (2023). "Obstetric Unit Closures and Racial/Ethnic Disparity in Health". In: *NBER Working Paper* w30986.

Choné, Philippe, Elise Coudin, and Anna Pla (2020). "Médecins en secteur 2 : les dépassements d'honoraires diminuent quand la concurrence s'accroît". In: *DREES* Études et Résultats 1137.
Choné, Philippe and Ching-to Albert Ma (2011). "Optimal health care contract under physician agency". In: *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pp. 229–256.
Choné, Philippe and Lionel Wilner (2022). "Financial Incentives and Competitive Pressure: The Case of the Hospital Industry". In: *Journal of the European Economic Association* 20.2, pp. 626–666.
Chopin, Nicolas and Omiros Papaspiliopoulos (2020). *An introduction to sequential Monte Carlo*. Springer Series in Statistics. Springer, Cham, pp. xii+378.
Cooper, Zack, Stephen Gibbons, Simon Jones, and Alistair McGuire (2011). "Does hospital competition save lives? Evidence from the English NHS patient choice reforms". In: *The Economic Journal* 121.554, F228–F260.
Currie, Janet M and W Bentley MacLeod (2020). "Understanding doctor decision making: The case of depression treatment". In: *Econometrica* 88.3, pp. 847–878.
Cutler, David, Jonathan S Skinner, Ariel Dora Stern, and David Wennberg (2019). "Physician beliefs and patient preferences: a new look at regional variation in health care spending". In: *American Economic Journal: Economic Policy* 11.1, pp. 192–221.
Dafny, Leemore S (2005). "How do hospitals respond to price changes?" In: *American Economic Review* 95.5, pp. 1525–1547.
Dash, Sabyasachi, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik (2019). "Big data in healthcare: management, analysis and future prospects". In: *Journal of Big Data* 6.1, pp. 1–25.
De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". In: *American Economic Review* 110.9, pp. 2964–2996.
Dubois, Pierre and Laura Lasio (2018). "Identifying industry margins with price constraints: Structural estimation on pharmaceuticals". In: *American Economic Review* 108.12, pp. 3685–3724.
Duchaine, Fanny, Guillaume Chevillard, and Julien Mousquès (2022). "Inégalités territoriales de répartition des infirmières libérales : quel impact des restrictions à l'installation en zones sur-denses et des incitations financières en zones sous-denses?" In: *IRDES* Questions d'économie de la Santé 87.
Eliason, Marcus, Lena Hensvik, Francis Kramarz, and Oskar Nordström Skans (2022). "Social connections and the sorting of workers to firms". In: *Journal of Econometrics*.
Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams (2016). "Sources of geographic variation in health care: Evidence from patient migration". In: *The quarterly journal of economics* 131.4, pp. 1681–1726.
— (2021). "Place-based drivers of mortality: Evidence from migration". In: *American Economic Review* 111.8, pp. 2697–2735.
Gao, Wayne Yuan, Ming Li, and Sheng Xu (2022). "Logical differencing in dyadic network formation models with nontransferable utilities". In: *Journal of Econometrics*.
Gaynor, Martin, Rodrigo Moreno-Serra, and Carol Propper (2013). "Death by market power: reform, competition, and patient outcomes in the National Health Service". In: *American Economic Journal: Economic Policy* 5.4, pp. 134–166.
Gottlieb, Daniel J, Weiping Zhou, Yunjie Song, Kathryn Gilman Andrews, Jonathan S Skinner, and Jason M Sutherland (2010). "Prices don't drive regional Medicare spending variations". In: *Health Affairs* 29.3, pp. 537–543.
Graham, Bryan S (2017). "An econometric model of network formation with degree heterogeneity". In: *Econometrica* 85.4, pp. 1033–1063.
Granger, Clive WJ (1969). "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society*, pp. 424–438.
Grünewälder, Steffen and Azadeh Khaleghi (2019). "Approximations of the restless bandit problem". In: *The Journal of Machine Learning Research* 20.1, pp. 514–550.
— (2021). "Oblivious data for fairness with kernels". In: *The Journal of Machine Learning Research* 22.1, pp. 9441–9476.
HCAAM (2022). *Organisation des Soins de proximité : Garantir l'accès de tous à des soins de qualité*. Rapport du Haut Conseil pour l'avenir de l'Assurance maladie.
Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.
Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul (1999). "An introduction to variational methods for graphical models". In: *Machine learning* 37, pp. 183–233.
Kessler, Daniel P and Mark B McClellan (2000). "Is hospital competition socially wasteful?" In: *The Quarterly Journal of Economics* 115.2, pp. 577–615.
Khaleghi, Azadeh and Gábor Lugosi (2023). "Inferring the mixing properties of a stationary ergodic process from a single sample-path". In: *IEEE Transactions on Information Theory*.
Khaleghi, Azadeh and Daniil Ryabko (2012). "Locating changes in highly dependent data with unknown number of change points". In: *Advances in Neural Information Processing Systems* 25.
— (2014). "Asymptotically consistent estimation of the number of change points in highly dependent time series". In: pp. 539–547.
— (2016). "Nonparametric multiple change point estimation in highly dependent time series". In: *Theoretical Computer Science* 620, pp. 119–133.
— (2020). "Clustering piecewise stationary processes". In: pp. 2753–2758.
Khaleghi, Azadeh, Daniil Ryabko, Jeremie Mari, and Philippe Preux (2016). "Consistent algorithms for clustering time series". In: *Journal of Machine Learning Research* 17.3, pp. 1–32.
Kraus, Johann M, Ludwig Lausser, Peter Kuhn, Franz Jobst, Michaela Bock, Carolin Halanke, Michael Hummel, Peter Heuschmann, and Hans A Kestler (2018). "Big data and precision medicine: challenges and strategies with healthcare data". In: *International Journal of Data Science and Analytics* 6, pp. 241–249.
Linde, Sebastian (2019). "The formation of physician patient sharing networks in medicare: Exploring the effect of hospital affiliation". In: *Health economics* 28.12, pp. 1435–1448.
Lucas-Gabrielli, Véronique, Catherine Mangeney, Fanny Duchaine, Laure Com-Ruelle, Abdoulaye Gueye, and Denis Raynaud (2022). "Inégalités spatiales d'accessibilité aux médecins spécialistes". In: *IRDES* Document de travail 87.
Luo, Jake, Min Wu, Deepika Gopukumar, and Yiqing Zhao (2016). "Big data application in biomedical research and health care: a literature review". In: *Biomedical informatics insights* 8, BII–S31559.
Luo, Jun (2014). "Integrating the huff model and floating catchment area methods to analyze spatial access to healthcare services". In: *Transactions in GIS* 18.3, pp. 436–448.
Luo, W. and Y. Qi (2009). "An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians". In: *Health and Place* 15.4, pp. 1100–1107.
Mahase, Elisabeth (2019). *Women in India face "extensive gender discrimination" in healthcare access*.
Mayor, Susan (2018). *Poor access to GP services is not linked to frequent A&E visits, finds study*.
Meyn, Sean (2022). *Control systems and reinforcement learning*. Cambridge University Press.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.
Mnih, Andriy and Russ R Salakhutdinov (2008). "Probabilistic matrix factorization". In: *Advances in neural information processing systems*, pp. 1257–1264.

Ni, Xin, Zhe Li, Xinping Li, Xiao Zhang, Guoliang Bai, Yingying Liu, Rongshou Zheng, Yawei Zhang, Xin Xu, Yuanhu Liu, et al. (2022). "Socioeconomic inequalities in cancer incidence and access to health services among children and adolescents in China: a cross-sectional study". In: *The Lancet* 400.10357, pp. 1020–1032.

Pastorino, Roberta, Corrado De Vito, Giuseppe Migliara, Katrin Glocker, Ilona Binenbaum, Walter Ricciardi, and Stefania Boccia (2019). "Benefits and challenges of Big Data in healthcare: an overview of the European initiatives". In: *European journal of public health* 29.Supplement_3, pp. 23–27.

PL., Delamater (2013). "Spatial accessibility in suboptimally configured health care systems: a modified two-step floating catchment area (M2SFCA) metric". In: *Health Place*.

Propper, Carol and John Van Reenen (2010). "Can pay regulation kill? Panel data evidence on the effect of labor markets on hospital performance". In: *Journal of Political Economy* 118.2, pp. 222–273.

Rozemberczki, Benedek, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar (2022). *The Shapley Value in Machine Learning*.

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton (2007). "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th international conference on Machine learning*, pp. 791–798.

Shan, Hanhuai and Arindam Banerjee (2010). "Generalized probabilistic matrix factorizations for collaborative filtering". In: *2010 IEEE International Conference on Data Mining*. IEEE, pp. 1025–1030.

Shen, Yu-Chu (2003). "The effect of financial pressure on the quality of care in hospitals". In: *Journal of health economics* 22.2, pp. 243–269.

Silhol (2020). "Pratiques des médecins généralistes dans les territoires devenus zones d'intervention prioritaire". In: *Insee Analyses* 51.

Tay, Abigail (2003). "Assessing competition in hospital care markets: the importance of accounting for quality differentiation". In: *RAND Journal of Economics*, pp. 786–814.

Torregrossa, François, Robin Allesiardo, Vincent Claveau, Nihel Kooli, and Guillaume Gravier (2021). "A survey on training and evaluation of word embeddings". In: *Int. J. Data Sci. Anal.* 11.2, pp. 85–103. DOI: 10.1007/s41060-021-00242-8. URL: https://doi.org/10.1007/s41060-021-00242-8.

Wan, Neng, Bin Zou, and Troy Sternberg (2012). "A three-step floating catchment area method for analyzing spatial access to health services". In: *International Journal of Geographical Information Science* 26.6, pp. 1073–1089.

Wu, Haixia, Chunyao Song, Yao Ge, and Tingjian Ge (2022). "Link Prediction on Complex Networks: An Experimental Survey". In: *Data Sci. Eng.* 7.3, pp. 253–278. DOI: 10.1007/s41019-022-00188-2. URL: https://doi.org/10.1007/s41019-022-00188-2.

Wu, Shiwen, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui (2022). "Graph Neural Networks in Recommender Systems: A Survey". In: 55.5.

Zeltzer, Dan (2020). "Gender homophily in referral networks: Consequences for the medicare physician earnings gap". In: *American Economic Journal: Applied Economics* 12.2, pp. 169–197.