Data Science Platform Applied to Health in Contribution to the Brazilian Unified Health System

Marcel Pedroso[1,*], Rebecca Salles[1,2,*], Raphael Saldanha[5], Vinicius Kreischer de Almeida[1], Gabriel Souto[1,6], Balthazar Paixão[2], Sérgio Ricardo de Borba Cruz[1], Carlos Cardoso[1,3], Victor Ribeiro[1,3], Raquel Gritz[1], Carmen Bonifácio[1], Matheus Miloski[1], Carlos Augusto de Sousa[7], Gizelton Pereira Alencar[8], Ariane Alves[1], Nelson Niero Neto[1], Letícia Sabbadini[1], Eduardo Ogasawara[2], Christovam Barcellos[4], Fabio Porto[3], Lucas Zinato Carraro[1] and Jefferson Lima[1,*]

[1]*Data Science Platform applied to Health (PCDaS)/Lis/Icict, Oswaldo Cruz Foundation (Fiocruz), Brazil*
[2]*Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ), Brazil*
[3]*National Laboratory for Scientific Computing (LNCC), Brazil*
[4]*Laboratory of Health Information (LIS)/Icict, Oswaldo Cruz Foundation (Fiocruz), Brazil*
[5]*National Institute for Research in Digital Science and Technology (INRIA), France*
[6]*PESC/COPPE,Federal University of Rio de Janeiro (UFRJ), Brazil*
[7]*DTIES/FCM, Rio de Janeiro State University (UERJ), Brazil*
[8]*School of Public Health, University of São Paulo (USP), Brazil*

Abstract

The Data Science Platform Applied to Health (PCDaS) is a research and technological development project that aims to develop and apply novel data analysis methods to public health data. It fills a technological gap between the variety of data sources available in legacy and unstandardized formats and the current needs and possibilities of Data Science applications to consume and explore data for the benefit of the Brazilian Health System. PCDaS provides democratic access to health-related datasets and information by requiring fewer technological abilities from its users while maintaining a continuously updated stack of technologies. As a data ecosystem, our primary goal is to provide secure and remote access to health data, technological tools, and a robust infrastructure provided by our platform to process and analyze a large amount of data that generally demand computational power often unavailable to researchers. The infrastructure consists of multi-region on-premise and cloud servers prepared to deal with the heavy analysis of Big Data from anywhere from multiple users simultaneously. Providing secure and remote access to health databases, whether in their original form or processed, is a daily breakthrough for a public health researcher. Knowing that there is a place where they can access integrated data in a standard format makes the research process much more manageable. To ensure quality, our data engineering and governance teams process these data sources following a gold standard based on cross-tables provided by the Health Ministry (the TabNET system) and decoding the original variables into meaningful names provided by the sources. It is very relevant to emphasize the comprehensive documentation of metadata, attributes, and the ETL (Extract, Transform, Load) process for databases. Every part of these steps is described in detail on the PCDaS website, ensuring the comprehension and reproducibility of the process. These features ensure that PCDaS users can effectively leverage the platform's resources and capabilities, enabling them to conduct research, perform data analysis, and collaborate within a secure and supportive environment to contribute to the Brazilian Health System.

Keywords

Public Health, PaaS, Data Science, Data Ecosystem

1. Introduction

The Data Science Platform applied to Health ("Plataforma de Ciência de Dados aplicada à Saúde" – *PCDaS*) is a research and technological development project of the Laboratory of Health Information ("Laboratório de Informação em Saúde" - LIS) from the Institute of Scientific and Technological Communication and Information in Health ("Instituto de Comunicação e Informação Científica e Tecnológica em Saúde" - ICICT), of the Oswaldo Cruz Foundation ("Fundação Oswaldo Cruz" - Fiocruz), in partnership with the Data Extreme Lab (DEXL) from the National Laboratory for Scientific Computing ("Lab-

Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — Workshop on Data Ecosystems (DEco'23), August 28 - September 1, 2023, Vancouver, Canada

*Corresponding author.

marcel.pedroso@icict.fiocruz.br (M. Pedroso); rebeccapsalles@acm.org (R. Salles); jefferson.lima@icict.fiocruz.br (J. Lima)
0000-0002-7323-2107 (M. Pedroso); 0000-0002-1825-0097 (R. Salles); 0000-0002-7560-1068 (J. Lima)

oratório Nacional de Computação Científica" - LNCC). Both Fiocruz and LNCC are two of the main research institutions in Brazil. Fiocruz is the most prominent institution of science and technology in health in Latin America, while LNCC is the first Brazilian institution in the field of Scientific Computing having the most powerful IT resources in Latin America.

PCDaS aims to develop and apply novel methods of data analysis on public health data, filling a technological gap between the variety of data sources of interest to the Public Health available in legacy and unstandardized formats and the current needs and possibilities of Data Science applications to consume and explore data for the benefit of the Brazilian Health System.

Data science emerged as an interdisciplinary field, covering the study of structured and non-structured data of different volume, complexity, variety, and other properties, also called "Big Data" [1]. Public Health may benefit from Data Science techniques and paradigms, with conditions to advance statistical and epidemiological practical applications to more complex and heterodox data sources.

Like other research fields, Public Health has historically been guided by a theory-driven or hypothesis-driven approach to science, with *a priori* assumptions. The new challenges imposed by Big-Data go beyond scaling-up computer servers to use the same methods. New approaches are needed to overhaul the methods to use better the full potential of available data in its diversity and complexity, leading to a data-driven approach to science [2, 3]. For the context of *PCDaS*, Data Science is interpreted as a field of study that can aid the discovery of knowledge of useful information from big or complex databases and aid the decision-making guided by data [4].

The Brazilian Health System is publicly funded and offers universal free health care coverage to the Brazilian population, known as the Unified Health System ("Sistema Único de Saúde" – SUS). It was established in 1988, along with the re-democratization process of the country. A component of the SUS is the Department of Informatics (DataSUS), which is responsible for gathering, organizing, and disseminating Brazilian health data.

The health data at DataSUS is structured by several Health Information Systems (HIS) dedicated to covering different aspects of a person's life cycle, resulting in specific HIS like the SINASC for birth data, SIH for hospital admissions data, and SIM for mortality data. There are other dozens of HIS maintained by the DataSUS, covering aspects such as vaccinations, ambulatorial services, records of health professionals, health services, health equipment, suspected cases of transmissible diseases, violence, and other themes. The anonymized raw data from those systems are publicly available through the DataSUS website.

In common, those HIS were created in different moments, trying to respond to different needs from epidemiological and administrative points of view. Most of those systems were implemented using 1990 legacy database technologies, such as dBase and other DOS and early versions of Windows applications. Its technological modernization is planned but taking place slowly by facing the challenges of covering a country with continental extensions with inequalities in Internet access, technical working force availability, political interference, and funding.

The DataSUS made Brazilian health data publicly available since its creation, fulfilling a mission of data dissemination, being a pioneer government agency of open data in the 90's. Open data principles implemented on HIS and the citizen rights to information access are keystones for the *PCDaS* creation.

Using the available data from DataSUS and other sources of relevant sociodemographic information, such as Census and population thematic inquiries, usually imposes steps of downloading and handling large amounts of files with storage needs and processing and filtering the data for specific research needs. Due to the nature of the available data formats in its sources, this Extraction-Transform-Load (ETL) process includes the use of legacy software, undocumented sources, and handling of larger-than-memory data by non-technological savvy users, such as social scientists, epidemiologists, geographers, and others.

The current ecosystem of data sources of health data and relevant sources of information for Public Health imposes on researchers and managers a long learning curve, with non-standardized and shared practices that lead to repeating the ETL processes among several research groups to obtain similar but not comparable results.

Aiming at providing a Data Science platform for Public Health, the *PCDaS* is structured as Platform-as-a-Service (PaaS) to cover aspects such as ETL, data analysis, data visualization, modeling, artificial intelligence, and knowledge dissemination. At PCDaS, we strive to create a community of data scientists who collaborate with SUS to offer advanced technology and scientific computing services. Our primary focus is to help manage, store, analyze, visualize, and share extensive data related to healthcare and its socio-environmental influences. Our services cater to researchers, professors, students, educational and research institutions, as well as government officials. Our objective is to advocate for positive advancements in public health policies and society as a whole.

Besides this introduction, Section 2 presents a literature review, a background on PaaS, and discusses related works. Section 3 further details the components of the *PCDaS* data ecosystem. Section 4 shares some of the main research projects furthered by *PCDaS* as well as their products and scientific results. Finally, Section 5 concludes and describes the plans for expansion and continuous improvement of *PCDaS*.

2. Literature review

Over the past years, industry and academia have shown a growing interest in Big Data and analytics. Despite advancements in computer systems, handling large-scale data remains an important challenge. Commonly, we encounter hardware and time limitations, which drive the improvement of data processing methods. Big Data refers to the vast amount of data created and exchanged. Its applications are characterized by the "3Vs": volume (information volume), velocity (data generation and consumption time), and variety (heterogeneous data sources) [5]. Additional dimensions such as veracity, validity, value, variability, venue, vocabulary, and vagueness have been proposed to complement the understanding of Big Data [1].

Over the past decade, data science has emerged as a highly relevant field. It encompasses a multidisciplinary approach, gathering different knowledge areas [6]. The fundamental objective of data science is to extract valuable insights and knowledge from data, leveraging analytical techniques and computational tools. Collaborative platforms such as Kaggle and OpenML performs a relevant role as players on data science community. For instance, the metioned platforms provides datasets, data documentation and some exploratory analysis.

In the context of public health, data science plays a crucial role. Goldsmith *et al.* [7] defines data science in public health as a discipline that focuses on formulating and answering questions related to public health and well-being through data-centric approaches. In recent years, public health researchers in Brazil have increasingly turned to data science tools and methodologies to understand the Brazilian health system better and improve healthcare outcomes. The analyses using the information provided by DataSus present themselves as a typical Big Data problem due to its volume.

DataSUS is the primary information system responsible for providing computational support to all instances of SUS. Despite its crucial role, accessing data from DataSUS can be challenging due to the fragmented nature of the available data sources, which makes it difficult for researchers and stakeholders to obtain a comprehensive view of the public health data landscape.

According to [8] *Platform-as-a-Service* (PaaS) is an interface that provides access to a complex set of technological components. Today, *PaaS* plays a central role in internet applications, abstracting architecture and infrastructure complexities for users. PaaS exhibits three significant characteristics [9]: it is internet resource centered, is open to third-party developers, and utilizes web API (Application Programming Interface).

PaaS offers access to its resources through network connections, delivering services to users. The concept of an open platform relies on enabling third-party developers to create internet-based value-added applications. This integration is facilitated through Application Programming Interfaces (APIs), granting third-party developers access to resources and services. Consequently, this accessibility yields tangible benefits for organizations, developers, and users.

One plausible solution to address Big Data challenges is through data ecosystems. Section 3 presents a more accurate description of our architecture, highlighting how we provide *PCDaS* as a *Platform-as-a-Service*. In this context, we have implemented a data ecosystem that facilitates collaborative efforts to enhance our understanding of public health in Brazil.

Data ecosystems transcend traditional production chains by incorporating three key characteristics: network, platform, and co-evolution [10]. As the mentioned work describes, networks within data ecosystems are formed by developers, providers, technology suppliers, and infrastructure. Platforms serve as the means through which ecosystem participants interact. Moreover, data ecosystems provide resources that enable participants to evolve through interactions among stakeholders and across different knowledge fields. A more specific discussion [8] about *PaaS* and the relation between their components. A data ecosystem can also be understood as a complex set of interactions between heterogeneous agents and their environment, similar to a biological ecosystem [11].

The increasing variety and volume of data have presented us with numerous challenges. Some solutions have been proposed to overcome time and hardware restrictions. For integrating data from heterogeneous sources while ensuring data quality, Ramalli et al. [12] propose using *SciExpeM*, a framework designed to speed up and support the development of scientific models. This work was further extended [13], where a data ecosystem was proposed specifically for chemical engineering.

The increasing adoption of IoT sensors has led to the continuous monitoring of daily activities. In Yu et al. [14], a data ecosystem is proposed to address predictive maintenance problems in an industrial context. In materials science, Blaiszik et al. [15] emphasize enhancing data ecosystems to develop new technologies. The paper presents two projects supporting machine learning applications in materials science.

Although the mentioned works primarily focus on engineering applications, data ecosystems can be applied across various domains. For instance, in an analysis of the effects of territorial politics and metropolitan governance, Kitchin and Moore-Cherry [16] discuss how fragmented governance can reduce economies of scale

Table 1
Summary of the related work

Platform	ETL and Documentation	Educational Services	Quality	Teams	Centralized View	Funding	Domain
PCDaS	✓	✓	✓	✓	✓	Mixed	Public Health
Kaggle	✓	✓	✓			Private	ML Applications
OpenML	✓		✓			Public	ML Applications
Data.gov	✓					Public	Governmental Data
Canadian Inst. for health information	✓	✓	✓	✓		Public	Governmental Data
European Union Open Data Portal	✓	✓				Public	Governmental Data
World Bank Open Data	✓	✓				Public	Global Development
DATASUS	✓					Public	Public Health
Portal Brasileiro de Dados Abertos	✓					Public	Governmental Data

and limit the effectiveness of public policies. These papers contribute to the growing body of knowledge on data ecosystems. They typically present domain-specific problems and strategies for solving Big Data challenges. Table 1 provides a comprehensive list of data ecosystems, and to the best of the author's knowledge, *PCDaS* is the first data ecosystem specialized in public health in Brazil.

We selected some features for comparing some of the existing data ecosystems:

- **Educational Services:** indicates if the given platform provides training and technological literacy;
- **Teams:** the data ecosystem provider explicitly allocates teams for supporting data-driven projects;
- **ETL and Documentation:** indicates if the platform provides ETL process and data documentation;
- **Quality:** indicates if the given data ecosystem provides some data quality discussion;
- **Centralized View:** data provided after cleaning and enrichment process;
- **Funding:** indicates how data ecosystem are founded; and
- **Domain:** application area.

An important player in data ecosystems is the World Bank[1], an international organization dedicated to promoting equity and reducing poverty worldwide. Among its objectives is the support of countries in improving their statistical capacity through advisory services, project support, partnership management, and financial resources.

Despite beeing a more mature data ecosystem when copared to the other presented in Table 1, Canadian Institute for Healt Information (CIHI) does not provides a centralized access for data. The institution releases frequent reports and metrics wich can provide a wider comprehention on canadian public health landscape, and also offers training on how to access the information made available.

It is also important do mention Global Health Observatory (GHO) data repository, is an World Health Organization initiative that aims to provide health-related statistics for all 194 Member States of Unated Nations. They provide access to more than 1000 indicators via an API interface on collective health topics, i.e.: mortality and burden of diseases, immunization, malaria among others.

Notably, most of the listed data ecosystems do not provide data from a centralized view. By centralizing data access, *PCDaS* can guarantee the data science community with simpler usage. This decision allows us to enrich data, add value and enable more complex analysis. By releasing microdata, *PCDaS* offers more flexibility for the platform users. It is also worth to mention that, openning the choice of making our methodology public ensure reproducibility.

3. PCDaS

As discussed in the previous sections, the utilization of health data from Brazilian Health Information Systems in research endeavors presents numerous complex challenges that demand careful consideration. These challenges can manifest from various aspects, including intrinsic characteristics of the data itself and the necessary infrastructure for handling and analyzing the vast volumes of data available.

From a data perspective, several critical issues arise when dealing with health data, including the integration of data from diverse sources, data processing tasks such as cleaning and enrichment, the challenge of working with pre-aggregated data lacking granularity, non-standard file formats, managing the sheer volume of data, and ensuring appropriate data modeling for various research needs.

From infrastructure perspective, having a robust environment that can provide the necessary computational power to acquire, process, and analyze large volumes of data while ensuring privacy and security is crucial.

[1] World Bank provides capacitation over mentoring and funding programs

In addition to meeting the hardware requirements, the infrastructure must encompass the selection of suitable tools and services that can effectively support the research objectives. This infrastructure entails considering factors such as data storage, processing capabilities, scalability, and the ability to implement robust privacy and security measures.

PCDaS was created by establishing a robust data ecosystem that effectively addresses several of the aforementioned challenges. One of our primary focuses is to provide services that facilitate the acquisition and exploration of data by users. By promoting data integration, *PCDaS* aims to foster collaboration among different research groups, encouraging active participation in data sharing and promoting the reuse of valuable information. This collaborative approach can potentially enhance knowledge exchange, accelerate research projects, and ultimately contribute to data-driven healthcare decision-making.

The initial version of the platform, named *PCDaS* 1.0, became available in 2016 [17]. Despite its limitations, this release marked the realization of the concept of offering users a unified platform for accessing and analyzing vast amounts of HIS data. Subsequently, in 2019, *PCDaS* 1.5 was introduced [18], featuring enhancements to the user interface, comprehensive tutorials to aid platform utilization, and the addition of new and updated public datasets. *PCDaS* 2.0 was launched in 2021 [19], incorporating a range of improvements, such as Single Sign-On (SSO) functionality for users, seamless integration with Google Colab for notebooks and tutorials, and a broader selection of public datasets. More recently, a RESTful API was developed and released to users of partner projects (such as the ones presented in Section 4), with the aim of simplifying access to datasets hosted on the platform. In addition to enhancing the user experience, the platform's backend undergoes continuous evolution through the expansion and optimization of its infrastructure, architecture, and tools.

Currently, *PCDaS* has around 1,450 active users, supporting dozens of research and technological development projects from a variety of groups (academics and from the government), with particular interest in research and analysis on Public Health and socio-environmental determinants of health. The following subsections describe in detail the components of the *PCDaS* data ecosystem.

3.1. Infrastructure and Architecture

The infrastructure of the platform is hosted in two computational environments, in different geographical locations. The first is located at the Data Processing Center of the National Laboratory for Scientific Computation (LNCC), a renowned Brazilian research institution with a strong focus on high-performance computing. The second part of the infrastructure is located at the Data Processing Center of Oswaldo Cruz Foundation (Fiocruz), in Rio de Janeiro, which is widely recognized as one of the leading research centers in the field of public health. By leveraging the complementary nature of infrastructure and staff members of these two institutions, *PCDaS* can provide computational power and services capable of meeting the specific needs of internal engineering teams and users interested in using the platform.

Regarding availability, the infrastrucutre located at Fiocruz complies with the ABNT NBR 15247 (Safe Storage Environment with Fire Resistance Classification and Test Method) and NBR 60529 (Protection of Electrical Equipment) standards. Additionally, it has 2 uninterruptible power supplies (UPS) and its own power generator, ensuring the operation of the equipment 24 hours a day and providing protection against humidity, corrosive gases, magnetism, and high temperatures.

Furthermore, since the services are located in different places, any network or power related issues in one location only impact the services in that location, minimizing overall disruption to the platform's services. Additionally, having different teams responsible for maintaining the separate infrastructures and tools allows for a more specialized focus on specific concerns. This fine-grained specialization ensures that each team can efficiently address maintenance tasks and any issues that arise, contributing to the overall stability and reliability of the platform.

The platform prioritizes using free and open-source software (FOSS) for its tools and services. This approach offers several advantages in terms of infrastructure maintenance. FOSS solutions are cost-effective, providing a more affordable alternative than commercial options. Additionally, FOSS software is transparent, allowing for greater visibility into the code and ensuring the platform's operations are built on trustworthy foundations. Finally, the customizable nature of FOSS enables the platform to adapt and meet specific requirements efficiently. This way, the platform can function with minimal impact even when budget constraints arise.

A simplified overview of the infrastructure and architecture of the platform is presented in Figure 1. Despite the specification of current infrastructure, tools, and services, the platform follows a tool-agnostic and evolutionary approach, with planned updates and migration to new tools and infrastructure when favorable. It can be seen that the platform provides a set of services tailored to specific needs. Internal engineering teams of the platform dispose of tools for ETL jobs (Apache Airflow and Jupyter Hub), as well as complete access (write, read) to a Data Warehouse solution (ElasticSearch). Platform users' services are focused on providing computational power (JupyterHub and Google Colaboratory), data analysis tools (Kibana), and data consumption inter-

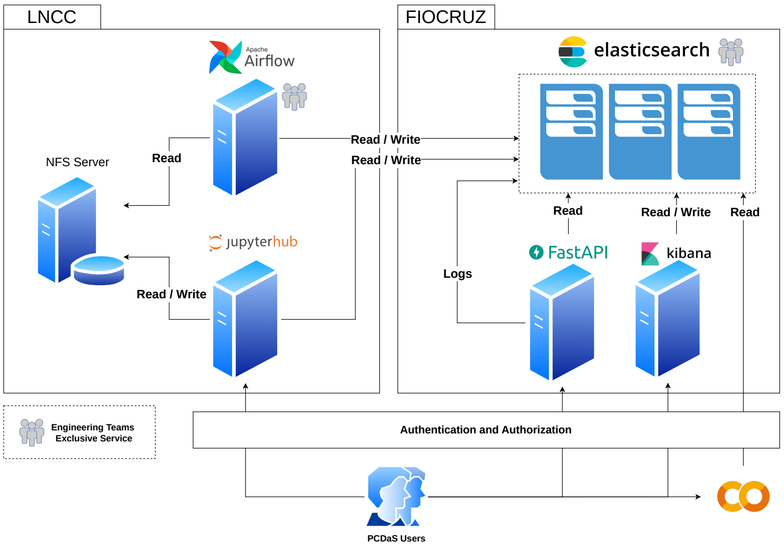


Figure 1: *PCDaS* Infrastructure and Architecture

faces (FastAPI).

Finally, it is important poiting out that, in addition to the mentioned infrastructure, *PCDaS* has the capability to harness cloud solutions when faced with scenarios that exceed the capacity or feasibility of our existing infrastructure and tools.

3.2. Services

As a data ecosystem, our main goal is to provide secure and remote access to health data, technological tools, and a robust infrastructure provided by our platform to process and analyze a large amount of data that researchers often lack the computational power to handle. This robust infrastructure consists of multi-region on-premise and cloud servers prepared to deal with the heavy analysis of Big Data from anywhere from multiple users simultaneously.

Providing secure and remote access to health databases available in *PCDaS*, whether in their original form or processed, is a daily breakthrough for a researcher. Knowing that there is a place where they can access integrated data in an ordinary format like CSV (comma-separated values) makes the research process much easier.

To ensure quality, our data engineering and governance teams process these data sources following a gold standard based on cross-tables provided by the Health Ministry in the TabNET system and decoding the original variables into meaningful names provided by the sources. By leveraging information provided by TabNET, we try to minimize potential issues regarding data inconsistencies by comparing the processed data against that found in TabNET.

It is very relevant to emphasize the comprehensive documentation of metadata, attributes, and the ETL (Extract, Transform, Load) process for databases. Every part of these steps is described in detail on the website, ensuring the comprehension and reproducibility of the process.

These features ensure that *PCDaS* users can effectively leverage the platform's resources and capabilities, enabling them to conduct research, perform data analysis, and collaborate within a secure and supportive environment.

As a way to organize the use and access to the platform, the users are organized into three categories: Basic Users, Academic Users, and Partner Users. What differentiates them are the services available for each one. Table 2 below summarizes the characteristics and features available to each user type in the *PCDaS* platform.

Basic Users have access to two key features: a) Support for data mining and predictive analysis through tutorials on Google Colab. Our team provides tutorials on how to utilize the data we have made available using open-source tools. b) Access to the community of researchers, data scientists, and *PCDaS* users in a public Slack Channel where they can interact and collaborate.

Academic Users have access to two additional levels of infrastructure, building upon the features available to Basic Users: c) Secure and reliable remote access to *PCDaS* via JupyterHub (Python or R), which ensures researchers have complete access and support while utilizing our technological infrastructure to prepare their analyses. d) Promotion of academic and partner projects on the *PCDaS* website. This feature provides a dedicated space

Table 2
Services by user category

Service	Basic Users	Academic Users	Partner Users
Support for data mining and predictive analysis through tutorials on Google Colab	✓	✓	✓
Access to the community of researchers, data scientists, and *PCDaS* users	✓	✓	✓
Secure and dedicated remote access to *PCDaS* via JupyterHub (Python or R)		✓	✓
Promotion of academic and partner projects on the *PCDaS* website		✓	✓
Basic training in Python or R			✓
Training of the research team for the use of *PCDaS*			✓
Extraction, Transformation, and Loading (ETL) of databases relevant to the research team's			✓

on our website to showcase the project's main goals and the individuals involved.

Finally, Partner Users receive the highest level of support and benefit from three additional features: e) Basic training in Python or R, offering classes on data usage, analysis, and manipulation with the chosen programming language. f) Training of the research team for using *PCDaS* to its fullest potential, and providing comprehensive support on leveraging our infrastructure and data sources. g) Extraction, Transformation, and Loading (ETL) of databases relevant to the research team's interests. Our experienced engineering team assists in extracting and structuring data, allowing researchers to focus on their project's core objectives without managing the technological aspects.

3.3. Data Management

This section outlines the Data Management process within *PCDaS*' data ecosystem. Data management refers to the activities undertaken to ensure the accuracy, integrity, and accessibility of the datasets. It encompasses data collection, organization, processing, storage, integration, and governance. Through effective data management practices, *PCDaS* strives to maintain the quality and reliability of data, enabling researchers to derive meaningful insights and make informed decisions based on reliable and comprehensive information.

3.3.1. Data Extraction

Data extraction is facilitated by utilizing two approaches: leveraging the orchestration capabilities of Apache Airflow or employing custom code on Jupyter Notebooks. Apache Airflow is employed when there is a need for automated data collection based on predefined schedules or events. On the other hand, Jupyter Notebooks are utilized for simpler data collection scenarios where a one-shot extraction is sufficient. This approach provides flexibility and ease of use for ad-hoc data collection tasks. By employing both Apache Airflow and Jupyter Notebooks, the platform can accommodate various data collection requirements and ensure efficient and effective data retrieval processes.

3.3.2. Data Transformation

Following data collection, the subsequent phase involves data processing. A significant challenge that users encounter when working with data from various HIS is the presence of diverse legacy and unstandardized file formats. Consequently, converting these files into user-friendly formats, such as CSV, Parquet, and other compatible file formats, is imperative. Although this task may appear straightforward, it often necessitates extensive research, understanding domain requirements, and potentially demanding the development of specialized libraries capable of handling these conversion processes effectively.

Another very important step of data transformation is data enriching. In the case of public HIS, data files often consist of numeric values representing categories or amounts, with separate files containing mappings for these categories to their respective string values. It is essential to perform data mappings that generate datasets containing numeric category values and their corresponding string representations. Such a process enhances the usability of the data. It enables easier data interpretation and analysis, ensuring meaningful insights can be derived from the enriched dataset.

In certain situations, it becomes necessary to integrate different datasets. This integration can be achieved by utilizing a shared column to enrich the resulting dataset with additional information. By joining datasets, valuable insights can be gained from the merged data, providing more comprehensive underlying information. This process enables researchers and analysts to leverage multiple datasets' combined knowledge and attributes, facilitating more informed analysis.

In cases where the final dataset is stored in our data warehouse, the data processing workflow includes the crucial step of data modeling. Data modeling ensures the data and the destination storage system's data model compatibility. It also ensures that the data is appropriately formatted to facilitate analytical queries. By performing data modeling, we ensure that the data is structured and organized to support efficient and effective analysis, enabling users to derive meaningful insights from the dataset.

3.3.3. Data Validation

Data validation is an essential step in the data processing workflows of HIS for several reasons. One of the main motivations is the reliance on manual human input for data entry in many of these systems. Despite verifications being performed in the source systems, it is not uncommon to encounter inconsistencies in the provided data. By conducting data validation, these inconsistencies can be identified and addressed, ensuring the accuracy and reliability of the processed data. Additionally, data validation plays a vital role in maximizing trust in the transformation process, as it helps prevent issues with the data from propagating to downstream tasks.

Data validation involves examining and verifying transformed data to ensure its adherence to predefined rules. These rules can be described in documents. It is the case of various HIS systems, as domain specialists can define them. Regardless of their origin, these rules serve as guidelines for ensuring data accuracy and consistency.

3.3.4. Data Load

Once the data has been transformed, the subsequent step involves loading the transformed data into the target systems. In the case of public datasets provided by the platform, this entails inserting the data into our ElasticSearch cluster and loading the transformed CSV files into a shared environment accessible to users. This load allows users to query the inserted data using ElasticSearch's powerful analytics engine. Additionally, users can utilize custom code on Jupyter or Colaboratory notebooks to perform more detailed analytics on the CSV files.

Another advantage of utilizing a distributed analytics system like ElasticSearch is its horizontal scalability. By consistently monitoring the system's resource usage, we can identify demand increases and, if needed, add additional machines to the cluster. This scalability feature ensures the platform can handle growing workloads effectively and maintain optimal performance.

The data can be loaded into their preferred systems in any acceptable format for private datasets generated for Partner Users. However, since the platform provides infrastructure that users may not possess, it is common for datasets to be loaded into a data warehouse. It allows centralized storage, efficient data management, and easy integration with the platform's analytical tools and services.

3.3.5. Data Access Management

Data acceess is managed by leveraging the concept of Elasticsearch roles. A role associates a collection of users to a set of permissions. These permissions can include different types of access (read, write) to specific indices, clusters, and objects in Elasticsearch and Kibana. Users are granted specific permissions they need to do their jobs, without giving them access to more data than they need. This helps to protect data from unauthorized access and allows to fine-tune the access that users have to different artifacts managed by the *PCDaS* plataform.

3.3.6. Example

Figure 2 illustrates the data management process for the System of Hospital Admissions (SIH) data. This workflow represents the data management process for single coverage of SIH, which releases data monthly.

The process starts with creating the folder structure to accommodate the downloaded and generated data. After this, a zip file containing the mappings used during the transformation step is downloaded from DataSUS FTP. The zip file decompression results in several files with CNV and DBF formats. These files represent the mappings and are parsed to CSV file format.

The next step is downloading data files representing each state's hospital admissions. As these files are in DBC file format, an additional step is necessary to convert them to a format compatible with popular processing tools. After format conversion, the dataset is validated by applying a set of rules provided in SIH documentation. After the pre-validation, data is enriched by mapping categorical values to respective strings and integrating them with other datasets. A post-validation step is performed after data enriching to verify the correctness of the data, and if the data is valid, it is loaded into our data warehouse.

The loaded data is validated against cross-tables available at the TabNET, a system provided by DataSUS, in which users can consult different data information about several HIS of SUS. Finally, the enriched data are gathered in a zip file and made available in a shared space to which platform users have access.

It is worth emphasizing that this process is designed specifically for SIH data, although it can be adapted for other datasets if needed. However, having distinct data management processes tailored to different datasets is more common.

3.4. Data Analysis and Visualization

As mentioned earlier, a primary objective of the platform is to facilitate users' access to HIS data. In order to achieve this objective, it is essential to collect, enhance, and make the data readily available and offer different interfaces

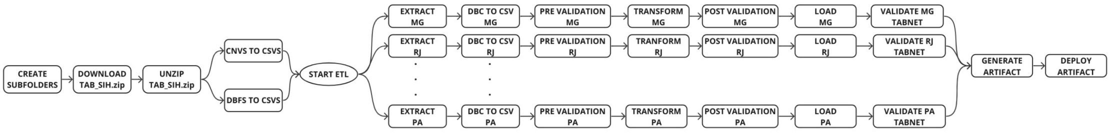


Figure 2: Data management example

that support distinct needs for analyzing and visualizing the information.

When client-side data manipulations are required, users can utilize our JupyterHub service or Google Colab notebooks to execute queries over our Data Warehouse or manipulate CSV or other formats made available by the platform. Additionally, an API developed using the FastAPI framework is provided to simplify the querying process by supporting SQL query language. Utilizing this API offers the advantage of logging user queries, enabling monitoring of request-related issues, and providing support when necessary.

Additionally, we offer a Kibana interface for users who prefer a more visual approach to data analysis or have limited familiarity with programming concepts. This interface provides a user-friendly and intuitive way to explore and analyze data stored in our ElasticSearch cluster, enhancing and simplifying the overall data analysis experience.

Examples of analysis of the platform's public data are available on our site, both in the form of Kibana dashboards and in tutorial examples that can be easily cloned and opened in Google Colaboratory's notebooks.

Various examples of analyses conducted on the platform's public data are presented on the *PCDaS* website. These examples are available in Kibana dashboards and tutorial examples that can be effortlessly cloned and opened in Google Colaboratory notebooks. These resources serve as valuable references for users looking to explore and learn from practical use cases of data analysis on our platform.

3.5. Data Transparency

Data transparency is crucial for enabling users and collaborators to comprehend the complete data generation process. By making the entire process transparent, there are several advantages. Firstly, users can verify that the process aligns with their expectations. They can suggest necessary modifications to meet their needs if any inconsistencies arise. This transparency empowers users to actively participate in shaping the data generation process, fostering a collaborative and inclusive environment.

For public datasets, to ensure data transparency, we offer detailed documentation that provides clear and comprehensive information about the entire data lifecycle, including data collection, processing, and utilization. This documentation is a valuable resource for users and collaborators, enabling them to understand how the data is obtained and handled.

Regarding private datasets, all code generated by the platform's engineering teams is shareable with our collaborators through simple notebooks and packaged applications. Additionally, we employ docker containers to facilitate the sharing process. This approach ensures seamless collaboration and enables our collaborators to work efficiently with the code and access the necessary data analysis and processing tools.

3.6. Data Protection

Following Brazilian guidelines on research that includes data from humans, especially from the Ethics Committee from the Escola Politécnica de Saúde Joaquim Venâncio, we adopt specific legal documents as the Data Use Commitment Term, or "Termo de Compromisso de Utilização de dados" (TCUD), where are stated the project leaders, objectives of the research, and how the data will be used and guarded. Projects that contain data with sensitive information follow specific guidelines from the Ethics Committee for data handling, storage, and retention policies.

3.7. Data FAIRness

Data sharing is a core element of *PCDaS* activities. Thus, with the aim of improving its quality and expanding the possibilities of reuse, it was decided to guide this process through the FAIR principles (Findable, Accessible, Interoperable, and Reusable).

Adopting FAIR principles makes it easier to discover and access data, while interoperability makes it possible to automate processes and integrate with other analytical tools, which can lead to new discoveries faster. In addition, there is the improvement of transparency and reliability, which can lead to greater reuse of data, in addition to promoting collaboration and innovation [13, 20, 21], which is fundamental to the field of public health.

At *PCDaS*, based on FAIR principles, information about data provenance, descriptive metadata, and the record of the data processing and dataset enrichment process are shared. The data is shared in open formats such as CSV and through an API, which makes the data readable by humans and machines, which together with the shared documentation, expands the possibility of interoperability and reuse.

4. Success cases

This section presents the main research projects conducted in partnership with *PCDaS* that provided services of scientific training, technological infrastructure, ETL, and data analysis and visualization. The described projects presented high impact especially in the field of public health as it can be observed by their achievements and public adoption of their products and scientific results. These projects exemplify, validate and represent the realization of the *PCDaS* goal of creating a community of data scientists and researchers, as well as government officials who collaborate with SUS through the use of advanced technology and scientific computing services, furthering positive advancements in public health policies and society as a whole.

4.1. GCE

Created in 2003 by the Bill & Melinda Gates Foundation, Grand Challenges Explorations is a program that invests in impactful research to solve significant health and development challenges worldwide [22]. Grand Challenges Brazil is the result of a partnership between the Department of Science and Technology (DECIT) of the Ministry of Health, the National Council for Scientific and Technological Development (CNPq), and the Bill & Melinda Gates Foundation. In Brazil, projects supported in data science [23] used innovative approaches to data analysis and modeling with the support of the *PCDaS* team in providing databases, infrastructure compatible with analysis in the Big Data scenario, teams dedicated to projects funded, and many others useful resources as can be seen in Table 3.

Brazilian Obstetric Observatory (OOBR) Among the main results achieved by *PCDaS* partner projects in the second call of the Grand Challenges Explorations – Brazil: Data Science to improve maternal and child health, women's and children's health in Brazil, it is worth mentioning that the efforts and contributions of OOBr (Brazilian Obstetric Observatory) [24], which monitors and analyzes the area of maternal and child health in Brazil. Remarkable advances have been achieved by the initiative in maternal, child, and obstetric health in a brief timeframe. Many resources have been generated, including articles, data visualization panels, indicators, and a book. Moreover, they have actively contributed to seminars and conferences, which caught the attention of researchers, journalists, and managers [25] [26] [27].

Accessibility to services and the reduction of maternal and child mortality (ASSMI) Another project with the support of the *PCDaS* team and infrastructure named "Accessibility to services and the reduction of maternal and child mortality" (ASSMI) focused on studying the impact of the displacements of many pregnant women when leaving their homes and cities and traveling long distances to health establishments that offer conditions for childbirth and neonatal care. Among the conclusions, the research indicated that the distances traveled for pregnant women to carry out deliveries and neonatal care increased in the period analyzed — between 2006 and 2017 and that infant deaths can be avoided by reducing the distances between mothers and the place of delivery or neonatal care [28]. A detailed summary of other project's contributions and support provided by *PCDaS* can be found at [29, 30]

Amplia Saúde The "Amplia Saúde" project [31], which also received support from *PCDaS*, offers an unparalleled alternative for visual analysis of maternal and child health data during the pre-and perinatal periods. The project takes into consideration environmental and climate factors. As part of this initiative, the First Big Data View Workshop in Maternal and Neonatal Health was developed during the second call of the Grand Challenges Explorations – Brazil. The workshop offers interactive tools for exploring and visually analyzing maternal and early neonatal health data correlated with data on environmental problems and extreme weather conditions.

Vax*Sim and MATRECI Immunizing children under five is highly effective and affordable in preventing dangerous and potentially fatal illnesses like polio, measles, diphtheria, tetanus, and pertussis. Sadly, around 20 million Brazilian children are still not receiving routine vaccination services, which creates disparities in healthcare coverage. With this scenario in mind, the project entitled "The Role of social media, the Bolsa-Família Program and Primary Health Care in immunization coverage for children under five in Brazil," — also called Vax*Sim, had the support of *PCDaS* in relevant stages of the process, such as monitoring the social network Twitter to analyze what was being published and commented by Brazilians on the topic of vaccination, use of data from the Information System on Live Births (SINASC) and the Information System on Mortality (SIM), in addition to sociodemographic data from municipalities, such as population and basic sanitation coverage [32].

Another project *PCDaS* supported is Breastfeeding in Brazil in the MATRECI Model: Mapping, Trend, Clus-

Table 3
Indicators of *PCDaS* usage for the second GCE call.

PCDaS Partnership	Resources allocated for GCE
Technological Infrastructure	- Dedicated servers at LNCC and Fiocruz; - Using the *PCDaS* API, over **40 tokens** were distributed, with over **1.6 million requests**; - The researchers had access to **8 Kibana Workspaces** containing over **10 Kibana dashboards**; - Researchers created over **290 Jupyter Notebooks** for data analysis and ETL, consisting of over **70 thousand lines of code**.
Communication and Management	- Efficient communication with **5,000 messages** exchanged through 14 Slack channels; - Task management in Workstreams.AI, with **10 Kanban boards** and over **200 tasks in motion**; - Operations for GCE were consolidated into one tribe consisting of **7 squads** and **17 professionals**.
Monitoring of projects	Over **100 meetings** with partners and **70 internal meetings** with dedicated squads were conducted.
Qualifications and training	- **3 introductory courses** to the use of *PCDaS*, and the R and Python language for data analysis; - Eight weeks of training, **18 synchronous classes** recorded on Youtube; - 54 entries: **21 for the R language and 33 for Python Courses**; - Five classes and **46 certified researchers**; - One year of access to **368 courses in data science** offered through Datacamp.

tering, and Impact. This project aims to track the implementation and progress of breastfeeding initiatives in primary health care (PHC). The team identified successful pro-breastfeeding programs and their impact. They created a dataset on PHC infrastructure, breastfeeding rates, and programs. They also qualified and verified SISVAN breastfeeding information [33]. These last two projects yielded significant contributions in published articles with their respective conclusions [34, 35, 36, 37, 38].

4.2. PNS

The National Health Survey ("Pesquisa Nacional de Saúde" - PNS) is a survey of the Brazilian Ministry of Health in partnership with the Brazilian Institute of Geography and Statistics (IBGE) focusing on the health conditions of the Brazilian population, obtained from a telephone questionnaire with population samples applied in distinct regions to provide subsidies to the formulation of public policies. *PCDaS* teamed up with other experienced researchers at Fiocruz to perform complex sampling calculations of 183 indicators referring to the surveys conducted in 2013 and 2019, make available the data referring to each stage of processing of these indicators, in addition to the metadata, the notebooks containing reproducible codes in R language and thematic dashboards, all open access.

These 183 indicators, organized into 15 modules and evaluated by statistical domains such as gender, region, age, color or race, household income, and level of education, seek to estimate eating habits, alcohol consumption, smoking, the practice of physical activities, chronic diseases, perception of states of health, older adult health, women's health, disability, access and use of health services and health care in the Brazilian population, among others.

As a result of the PNS project, a website was generated that gathers the previously mentioned information about the indicators in addition to dashboards with maps and figures, making it more accessible for researchers to carry out specific queries [2].

From its launch on April 1, 2021, to May 23, 2023, search and access to the site reached 50k users, and the engagement rate that measures user interactions on the site in terms of navigation was 51.5% distributed in Brazil (94.63%), United States (3.02%) and Portugal (0.41%). Table 4 presents the five most accessed pages.

Table 4
The five most accessed pages in the PNS site

Page	User access
PNS (Home)	36.59%
Indicators Panel	29.03%
Data Bases	6.82%
Requests base	4.99%
R program Notebooks	3.93%

The PNS website has been referenced in articles such as [39], a study about the elderly, socioeconomic factors and functionality, [40] surveillance of chronic diseases, [41] a study of the relationship between smoking and the worsening of Covid-19, at the congress as [42]. Other websites such as [43] are a precious resource for public health managers, thus contributing to the dissemination of information and support for new scientific collaborations.

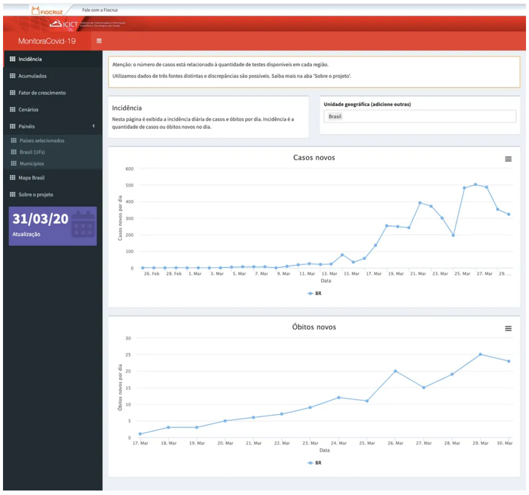
[2] The PNS project website is available at https://www.pns.icict.fiocruz.br/painel-de-indicadores-mobile-desktop/

4.3. MonitoraCovid-19

The MonitoraCovid-19 was created within the scope of ICICT/Fiocruz. Its main objective was to overcome the challenge of lacking open and centralized data from the Federal Government on COVID-19 cases. The project gathered data from different state governments to address this issue. Such data were consolidated into a single dataset and made available through platforms such as Kaggle and GitHub [44].

The project was crucial during many moments of the pandemic and served as the primary source of data for media outlets such as newspapers [45, 46, 47], news portals [48, 49] and also a Public Civil Action by the Federal Public Ministry [50]. Throughout this period, the *PCDaS* managed and maintained the infrastructure, ensuring its availability even when faced with thousands of simultaneous accesses following its dissemination in major media channels and social networks.

Figure 3 depicts the interactive content structure accessible to the public. The left black panel displays various sections, including Incidence, Accumulated Cases, Growth Factor, Scenarios, Dashboards (Brazil, States, and municipalities), an "About the Project" section, and the date of the last update. The central portion of the page showcases the plots corresponding to the user's selection. In the provided image, two graphs from the Brazil agency are displayed. The upper graph represents new cases per day, while the lower graph represents new daily deaths.



Atenção: o número de casos está relacionado à quantidade de testes disponíveis em cada região.

Utilizamos dados de três fontes distintas e discrepâncias são possíveis. Saiba mais na aba 'Sobre o projeto'.

Incidência

Nesta página é exibida a incidência diária de casos e óbitos por dia. Incidência é a quantidade de casos ou óbitos novos no dia.

Unidade geográfica (adicione outras): Brasil

Figure 3: The MonitoraCovid-19 website during pandemics [45]

The project's impressive achievements throughout its trajectory garnered significant attention and recognition. It recorded over 300,000 accesses and nearly 800,000 sessions, highlighting its widespread usage and impact. As a testament to its success, the project was honored with the second-place award in the Covid-19 Challenges organize by the National School of Public Administration (ENAP) [51].

4.4. SUS Ombudsman

In partnership with the "Instituto Aggeu Magalhães – Fiocruz Pernambuco" and with the Brazilian Health Ministry Ombudsman, the *PCDaS* supports a data science project that daily collects information about SUS health users' demands and its profiles.

This anonymized information is received in an automatic workflow, processed and stored at our ElasticSearch cluster, and visualized on Kibana dashboards. Currently, there are 4.993.669 user demands and 2.089.537 user profiles.

Those interactive data visualization panels are being used by the Health Ministry Ombudsman to monitor the number of requisitions, their topics, and user profile to better organized its actions and guide the Health Ministry's actions through its policies.

4.5. GeoAcess

The GeoAcess project, a partnership between Fiocruz and IBGE researchers, studies the geographical access to health services in Brazil by using a validated methodology and software (ACCESSMOD 5) developed by the World Health Organization. This project is funded by a research grant called "Fiocruz Inova: Geração de Conhecimento - Novos Talentos" from Fiocruz.

PCDaS provided its technical expertise to host the software execution and the necessary databases, which include massive raster data, helping the researchers to gain insights from the results with custom data visualizations.

4.6. Academic researches

Among the services provided by *PCDaS*, the platform also supports academic master's and doctoral projects. Building a network of partnerships, *PCDaS* provides infrastructure and assistance for academic research and provides a plan for academic users, whose services are described in Table 2.

Among the master and doctoral projects, to date, *PCDaS* has a total of 7 supported projects, which have generated dissertations and articles in the area of health [52, 53].

With graduate students from different educational institutions, such as the National School of Public Health Sérgio Arouca (ENSP), Federal Centers for Technological Education of Rio de Janeiro (CEFET/RJ), Graduate Program in Information and Communication in Health (PPGICS – ICICT/Fiocruz), Federal University of Rio de Janeiro (UFRJ) and National Laboratory of Scientific Computing

(LNCC). *PCDaS* provides support to students throughout Brazil designating a tutor (data scientist) to help with the tasks of ingestion and treatment of data from their research.

With themes such as: “COVID-19 Unifying Panel in the Favelas: territorialized analysis of morbidity and mortality associated with the new coronavirus in the municipality of Rio de Janeiro”; “Proposal for analysis of the scientific publication deposited in the Arca through automated processes”; and “Distance matters: Commuting networks for access to hospital childbirth in Brazil” the research projects were supported both in terms of making data available for study and in providing technological capacity to perform data analysis.

The master’s project “Discovering divergent association rules: A case study of Malaria in the Legal Amazon”, for example, used approximately 16GB of data from the infrastructure made available by *PCDaS* for the construction of analyses that resulted in the article [53]. The image 4 defines the percentage of samples confirming suspected malaria cases in some states of Brazil between the years 2009 to 2015.

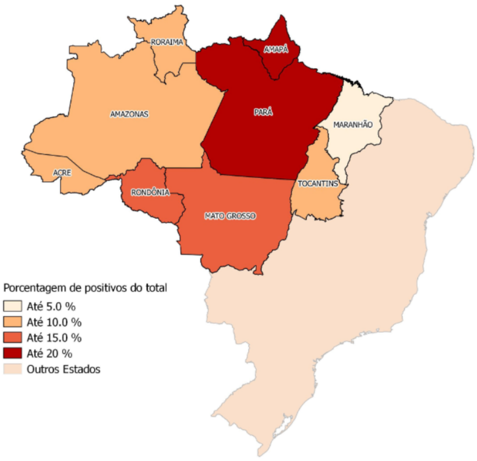


Figure 4: Results of analysis of the academic project "Discovering divergent association rules: a case study of Malaria in the Legal Amazon"

Another master’s project: “Basis Project: Study of tools, techniques, and methods that allow experimentation and understanding of data” used databases provided by the platform with information on mortality, births, and health facilities containing maternity hospitals for the construction of an environmental analysis with different information [54]. Data provided by *PCDaS* enabled the analysis of neonatal mortality by days of life of newborns in Brazil.

5. Conclusions

Adopting innovative technologies in research fields usually faces a dilemma between the gains and advantages of novel technologies against the trust and reliability recognized by well-known standards. *PCDaS* assist in solving this dilemma in Public Health research by providing a platform where its users and partners can focus more on its research questions and less on technological challenges by adopting validated and modern ready-to-use data and tools provided by the platform.

PCDaS helps to provide democratic access to health-related datasets and information by requiring less technological abilities from its users. We maintain a continuingly updated stack of technologies and schemes where the user does need to fully understand its details to use, but it is completely accessible and transparent to be trusted.

By adopting this strategy, *PCDaS* has grown a community of users and partners that takes advantage of using common tools and data validated by each other. This strategy proved to be valuable, being recognized by the projects awards and publications fostered by *PCDaS*. We can highlight the partnership with the “Open Knowledge Brasil” initiative at the “Querido Diário nas Universidades” project. This project aims to map the research ecosystem at universities and document the requirements to structure a workflow between academic projects and Brazilian FOSS initiatives.

PCDaS gained strength by adopting a sustainable partnership strategy, providing and maintaining a friendly, healthy, and productive work environment among our workforce and with our partners.

Another important initiative to foster a data scientists community in Public Health is the partnership established with DataCamp to provide premium access without costs to our members, students, and partners, aiming qualified training on learning tracks, especially on Data Science and Machine Learning, on Python and R languages.

The *PCDaS* team trained, in person and remotely, around 200 students, partners, and government managers by offering a public course called Data Science applied to Health, in its fifth edition, to classes of graduation, master’s, and doctorands.

To expand and empower our community of users and partners, there are plans to include more HIS data into the platform, as well other related sources of information relevant to Public Health, such as sociodemographic and environmental data. We plan to expand our partnerships to more national and international research groups and government agencies.

PCDaS created and maintains a data ecosystem adequate to a Universal Health System, facing the technological challenges to provide a platform that strengthens the SUS by aiding, with data and technological tools, the implementation and monitoring of health policies for the Brazilian population. To ensure the success of PCDaS, we have planned to (i) expand our institutional partnerships, (ii) provide ongoing training for health data scientists, (iii) make new datasets available, and (iv) continuously improve the software ecosystem that supports the Platform.

Acknowledgments

This work was supported by Institute for Scientic and Technological Communication and Information on Health (Icict/Fiocruz) in partnership with the National Laboratory of Scientific Computing (LNCC). Research Fund of the Rio de Janeiro State Research Foundation (FAPERJ), National Council for Scientific and Technological Development (CNPq), Bill and Melinda Gates Foundation, INOVA Fiocruz Program, Higher Education Personnel (CAPES). Various PCDaS partners provided additional funding. We also thank for contributing Artur Ziviani (in memoriam), João Vitor Maués Dias Carneiro, Jorge Nundes, Leandro Zirondi, Marcelo Rabaço, Marcus Vinicius Carneiro Magalhães, Matheus Ferreira. Special thanks to Igor da Silva Morais, who materialized the ideas and aspirations of the PCDaS, "live long and prosper"!

References

[1] K. Borne, Top 10 big data challenges a serious look at 10 big data v's, Blog Post 11 (2014).

[2] R. d. F. Saldanha, C. Barcellos, M. d. M. Pedroso, Ciência de dados e big data: o que isso significa para estudos populacionais e da saúde?, Cadernos Saúde Coletiva 29 (2021) 51–58. URL: https://doi.org/10.1590/1414-462X202199010305. doi:10.1590/1414-462X202199010305.

[3] S. A. Bohon, Demography in the big data revolution: Changing the culture to forge new frontiers, Population Research and Policy Review 37 (2018) 323–341. URL: http://link.springer.com/10.1007/s11113-018-9464-6. doi:10.1007/s11113-018-9464-6.

[4] PCDaS, Plataforma de ciência de dados aplicada à saúde, 2023. URL: https://pcdas.icict.fiocruz.br. doi:10.7303/syn25882127, laboratório de Informação em Saúde (Lis). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict). Fundação Oswaldo Cruz (Fiocruz).

[5] D. Laney, et al., 3d data management: Controlling data volume, velocity and variety, META group research note 6 (2001) 1.

[6] J. M. Wing, The data life cycle, Harvard Data Science Review 1 (2019) 6.

[7] J. Goldsmith, Y. Sun, L. Fried, J. Wing, G. W. Miller, K. Berhane, The emergence and future of public health data science, Public Health Reviews (2021) 4.

[8] D. Beimborn, T. Miletzki, S. Wenzel, Platform as a service (paas), Wirtschaftsinformatik 53 (2011) 371–375.

[9] C. Lv, Q. Li, Z. Lei, J. Peng, W. Zhang, T. Wang, Paas: A revolution for information technology platforms, in: 2010 International Conference on Educational and Network Technology, IEEE, 2010, pp. 346–349.

[10] M. I. S. Oliveira, B. F. Lóscio, What is a data ecosystem?, in: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, dg.o '18, Association for Computing Machinery, New York, NY, USA, 2018. URL: https://doi.org/10.1145/3209281.3209335. doi:10.1145/3209281.3209335.

[11] O. I. Dictionary, A new oxford dictionary (1962).

[12] E. Ramalli, G. Scalia, B. Pernici, A. Stagni, A. Cuoci, T. Faravelli, Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering, Frontiers in big Data 4 (2021) 663410.

[13] E. Ramalli, B. Pernici, et al., From a prototype to a data ecosystem for experimental data and predictive models, in: Proc. of the First International Workshop on Data Ecosystems (DEco'22), CEUR-WS, 2022, pp. 18–26.

[14] W. Yu, T. Dillon, F. Mostafa, W. Rahayu, Y. Liu, A global manufacturing big data ecosystem for fault detection in predictive maintenance, IEEE Transactions on Industrial Informatics 16 (2019) 183–192.

[15] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, I. Foster, A data ecosystem to support machine learning in materials science, MRS Communications 9 (2019) 1125–1133. doi:10.1557/mrc.2019.118.

[16] R. Kitchin, N. Moore-Cherry, Fragmented governance, the urban data ecosystem and smart city-regions: the case of metropolitan boston, Regional Studies 55 (2020) 1913.

[17] André Bezerra, Fiocruz disponibiliza plataforma de ciência de dados aplicada à saúde, 2016. URL: https://www.icict.fiocruz.br/content/fiocruz-disponibiliza-plataforma-de-ci%C3%AAncia-de-dados-aplicada-%C3%A0-sa%C3%BAde.

[18] André Bezerra, Lançada a versão 1.5 da plataforma de ciência de dados aplicada à saúde, 2019. URL: https://www.icict.fiocruz.br/content/lan%C3%A7ada-vers%C3%A3o-15-da-plataforma-de-ci%C3%AAncia-de-dados-aplicada-%C3%A0-sa%C3%BAde.

[19] Ariane Alves, Pcdas, a plataforma pública e gratuita de dados da saúde, ganha versão 2.0, 2021. URL: https://www.icict.fiocruz.br/content/pcdas-plata

forma-publica-e-gratuita-de-dados-da-saude-ganha-versao-20.

[20] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[21] J. Wise, A. G. de Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, G. Mellino, I. Harrow, I. Smith, J. Taubert, et al., Implementation and relevance of fair data principles in biopharmaceutical r&d, Drug discovery today 24 (2019) 933–938.

[22] G. C. Brazil, Quem somos, ????. URL: https://www.grandchallengesbrazil.org/quem-somos/.

[23] G. C. Brazil, Ciência de dados - chamadas, ????. URL: https://www.grandchallengesbrazil.org/chamada/ciencia-de-dados/.

[24] PCDaS, Observatório obstétrico brasileiro, 2022. URL: https://pcdas.icict.fiocruz.br/rede-de-parcerias/observatorio-obstetrico-brasileiro/.

[25] Fiocruz, Conheça o observatório obstétrico brasileiro covid-19 (oobr covid-19), 2021. URL: https://portaldeboaspraticas.iff.fiocruz.br/atencao-mulher/conheca-o-observatorio-obstetrico-brasileiro-covid-19-oobr-covid-19/.

[26] OOBr, Oobr srag, 2021. URL: https://observatorioobstetrico.shinyapps.io/covid_gesta_puerp_br/.

[27] PCDaS, Com o apoio da pcdas, disponibilização de dados e análises sobre mortalidade e morbidade das mães brasileiras são alvo de pesquisa, 2023. URL: https://pcdas.icict.fiocruz.br/publicacoes/com-o-apoio-da-pcdas-disponibilizacao-de-dados-e-analises-sobre-mortalidade-e-morbidade-das-maes-brasileiras-sao-alvo-de-pesquisa/.

[28] PCDaS, Pesquisa analisa o impacto das distâncias percorridas por gestantes brasileiras para realização do parto, 2023. URL: https://pcdas.icict.fiocruz.br/publicacoes/pesquisa-analisa-o-impacto-das-distancias-percorridas-por-gestantes-brasileiras-para-realizacao-do-parto/.

[29] PCDaS, Comunidade de analistas de dados utiliza a pcdas em oficina presencial, 2023. URL: https://pcdas.icict.fiocruz.br/publicacoes/comunidade-de-analistas-de-dados-utiliza-a-pcdas-em-oficina-presencial/.

[30] PCDaS, Pcdas encerra segundo ciclo de apoio ao gce/gates com avanços e conquistas, 2023. URL: https://pcdas.icict.fiocruz.br/publicacoes/pcdas-encerra-segundo-ciclo-de-apoio-ao-gce-gates-com-avancos-e-conquistas/.

[31] AmpliaSaude, Amplia saude site, 2021. URL: https://ampliasaude.org/en/.

[32] PCDaS, Vax*sim, 2021. URL: https://pcdas.icict.fiocruz.br/rede-de-parcerias/covac/.

[33] GCE, Breastfeeding in brazil in the matreci model: Mapping, trend, clustering and impact, 2021. URL: https://gcgh.grandchallenges.org/grant/breastfeeding-brazil-matreci-model-mapping-trend-clustering-and-impact.

[34] P. de Moraes Mello Boccolini, L. Baroni, L. de Almeida Relvas-Brandt, R. de Abreu Junqueira Gritz, et al., Brazilian spatial, demographic, and socioeconomic data from 1996 to 2020., BMC Research Notes 15 (2022) 159–159.

[35] R. F. S. Alves, C. S. Boccolini, L. R. Baroni, P. d. M. M. Boccolini, Primary health care coverage in brazil: a dataset from 1998 to 2020, BMC Research Notes 16 (2023) 63.

[36] P. d. M. M. Boccolini, C. S. Boccolini, L. de Almeida Relvas-Brandt, R. F. S. Alves, Dataset on child vaccination in brazil from 1996 to 2021, Scientific Data 10 (2023) 23.

[37] C. L. Szwarcwald, C. S. e. a. Boccolini, Covid-19 mortality in brazil, 2020-21: consequences of the pandemic inadequate management, Archives of Public Health 80 (2022) 255.

[38] G. R. e. a. Salles R, Ribeiro VPD, A comprehensive integrated dataset on brazilian health facilities: from 2005 to 2021., PREPRINT available at Research Square (2022) (????). doi:https://doi.org/10.21203/rs.3.rs-2358225/v1.

[39] R. J.P., G. Oliveira, N. J.C.D., B. A.M.G., B. Â.J.G, Impact of clinical and socio-economic factors and self-perception of health on the functionality of the elderly, Geriatr Gerontol Aging 11 (2017) 124–132. doi:https://doi.org/10.5327/Z2447-211520171700051.

[40] M. D.C., S. A. da., G. C.S., S. S.R., O. M. de., S. L.M.V., C. R.B., P. C.A., R.-N. E.L.G., Monitoring the goals of the plans for coping with chronic non-communicable diseases: results of the national health survey, brazil, 2013 and 2019, Epidemiologia e Serviços de Saúde 31(spe1) (2022) e2021364. doi:https://doi.org/10.1590/SS2237-9622202200008.especial.

[41] C. M. Peixer, T. R. Camargo, L. L. L. Silva, L. A. Colnago, L. L. Ferronatto, G. M. Lindenberg, O uso de tabaco e o desenvolvimento do covid-19 em adultos de 18 a 59 anos, uma breve revisão de literature / tobacco use and the development of covid-19 in adults aged 18 to 59 years, a brief literature review., Brazilian Journal of Development 8(3) (2022) 19226–19246. doi:https://doi.org/10.34117/bjdv8n3-253.

[42] A. MELO MENDONÇA, M.; SIQUEIRA ROCHA, Expectativa de vida com e sem doença crônica de coluna no brasil: Estudo comparativo a partir da pesquisa nacional de saúde, nos anos de 2013 a 2019., Refas - Revista Fatec Zona Sul 9, n. 2 (2022) 49–62.

doi:DOI:10.26853/Refas_ISSN-2359-182X_v09n02_08.

[43] A. Ogata, Rh pra você, 2022. URL: https://rhpravoce.com.br/colunistas/pns-um-recurso-precioso-para-os-gestores-em-saude-corporativa/.

[44] R. d. Saldanha, D. R. Xavier, M. d. Magalhães, P. R. Souza Junior, M. d. Pedroso, C. Barcellos, 14 - monitoracovid-19: Informação e disseminação de indicadores em uma pesquisa multidisciplinar, Covid-19 no Brasil: cenários epidemiológicos e vigilância em saúde (2021) 229–249. doi:10.7476/9786557081211.0015.

[45] F. Grandin, Ferramentas criadas por pesquisadores auxiliam no monitoramento da pandemia de coronavírus no brasil, 2020. URL: https://g1.globo.com/bemestar/coronavirus/noticia/2020/04/01/ferramentas-criadas-por-pesquisadores-auxiliam-no-monitoramento-da-pandemia-de-coronavirus-no-brasil.ghtml.

[46] CONASS, Monitoracovid-19: ferramenta online permite monitorar avanço da epidemia no brasil, dia a dia, 2020. URL: https://www.conass.org.br/monitoracovid-19-ferramenta-online-permite-monitorar-avanco-da-epidemia-no-brasil-dia-a-dia/.

[47] A. L. Azevedo, Sistema que monitora covid-19 mostra avanço do coronavírus no interior e põe em xeque redução do isolamento, 2020. URL: https://oglobo.globo.com/saude/coronavirus/sistema-que-monitora-covid-19-mostra-avanco-do-coronavirus-no-interior-poe-em-xeque-reducao-do-isolamento-24359218.

[48] A. Bezerra, Monitoracovid-19 avalia desigualdades no processo de vacinação, 2021. URL: https://agencia.fiocruz.br/monitoracovid-19-avalia-desigualdades-no-processo-de-vacinacao.

[49] Plox, Covid-19 ganha força com a vacinação estagnada, 2022. URL: https://plox.com.br/noticia/01/07/2022/covid-19-ganha-forca-com-a-vacinacao-estagnada.

[50] MPF, AÇÃo civil pÚblica com pedido de tutela provisÓria de urgÊncia, 2021. URL: https://www.mpf.mp.br/df/sala-de-imprensa/docs/copy2_of_ACP_DF.pdf.

[51] ENAP, Saíram os vencedores do desafios covid-19, 2020. URL: https://www.enap.gov.br/pt/acontece/noticias/sairam-os-vencedores-do-desafios-covid-19.

[52] Tarini de Souza Faria, Painel unificador covid-19 nas favelas: análise territorializada da morbimortalidade associada ao novo coronavírus no município do rio de janeiro, 2022. URL: https://pcdas.icict.fiocruz.br/wp-content/uploads/2022/09/Projeto-de-mestrado-Tarini-de-Souza-Faria.pdf.

[53] L. Baroni, B. Paixão, A. Chrispino, G. Guedes, C. Barcellos, M. Pedroso, E. Ogasawara, Análise exploratória da malária na amazônia brasileira por meio da plataforma de ciência de dados aplicada à saúde, in: Anais do XIII Brazilian e-Science Workshop, SBC, Porto Alegre, RS, Brasil, 2019. URL: https://sol.sbc.org.br/index.php/bresci/article/view/10025. doi:10.5753/bresci.2019.10025.

[54] Letícia Ange Pozza, Projeto basis: Estudo de ferramentas, técnicas e métodos que possibilitem a experimentação e entendimento dos dados, 2022. URL: https://dot.theoddstudio.com.br/.