

Article

A Cloud-Based Multidimensional Big Data Framework for Healthcare Analytics: Bridging OLAP and Cognitive Insights in Chronic Pain Management[†]

Alfredo Cuzzocrea^{1,2,*,†}  and Abderraouf Hafsaoui¹ ¹ iDEA Lab, University of Calabria, 87036 Rende, Italy; abderraouf.hafsaoui@unical.it² Department of Computer Science LIPADE, University of Paris City, 75013 Paris, France

* Correspondence: alfredo.cuzzocrea@unical.it

[†] This paper is an extension of the paper: Cuzzocrea, A.; Hafsaoui, A.; Gallo, C. From Theory to Practice of Multidimensional Big Data Analytics Over Big Healthcare Data: A Real-Life Case Study. In Proceedings of the 28th International Database Engineered Applications Symposium, Bayonne, France, 26–29 August 2024, pp. 156–171. https://doi.org/10.1007/978-3-031-83472-1_11.[‡] This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France.

Abstract

By considering the real-life research project *Pain-RELife*, which focuses attention on *big data management and analytics tools* over patients suffering from chronic pain, located in the Region Lombardy of North Italy, this paper provides a relevant journey from theory to practice about so-called *Multidimensional Big Data Analytics tools* over (real-life) *big healthcare datasets*, as dictated by the effective project goals. This constitutes an effective contribution to the state-of-the-art research. Our conceptual and theoretical results are corroborated by a comprehensive campaign of real-life experimental results focused on advanced tools such as *Multidimensional Clustering* and *Multidimensional Regression*.

Keywords: multidimensional big data analytics tools; healthcare data science; advanced big data systems

1. Introduction

Starting from the results of the real-life national project POR 2014–2020 Regione Lombardia “Pain-RELife—Ecosistema big data sostenibile e integrato per la continuità della cura e il supporto alla decisione dei pazienti con dolore”—“Pain-RELife—Sustainable and integrated big data ecosystem for continuity of care and decision support for patients with pain” (Project ID: 1173269), this paper reports on the practical and experimental results we obtained, thanks to the application of the well-known *Multidimensional Big Data Analytics paradigm* (e.g., [1–4]), which represents the foundational theory of this paper, along with lessons learned.

In more detail, in the *Pain-RELife* project, starting from a *kernel* real-life *chronical-pain-patient dataset*, with a special focus on important diseases, such as *Fibromyalgia*, *Ictus*, and *Cancer*, we *artificially* build a *synthetic multidimensional database* named *PainRELifeDB* via ad hoc *Random Sampling Techniques* (e.g., [5]) that strictly adhere to the multidimensional nature of the schema. One specific characteristic of the *PainRELifeDB* database is represented by the *pain-level questionnaire*, where patients report on their levels of pain following the therapy. This questionnaire is completely encoded in the database and, as a consequence, in the multidimensional big data analytics tools developed by our *Cloud-based framework*,



Academic Editor: Isaac Triguero

Received: 28 May 2026

Revised: 23 July 2026

Accepted: 27 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

as a relevant extension of classical big data analytics tools (e.g., [6]). As an immediate consequence, the knowledge derived from such multidimensional analysis is of primary interest for the underlying healthcare analytics (e.g., [7]).

Inspired by this line of research, this paper proposes an overview of our *Cloud-Based Multidimensional Big Data Analytics framework*, along with the tools we introduce and experiment in the context of this new big data analytics methodology, namely *Multidimensional Clustering* (e.g., [8]) and *Multidimensional Regression* (e.g., [9]), which are applied to the specific case of *big healthcare data analytics* (e.g., [10,11]). The first one is a powerful extension of classical clustering tools directly applied to the *core* multidimensional modeling of the target dataset via meaningful dimensions and measures. The second one is a nice generalization of the one-dimensional regression technique along all the *simultaneous* dimensions of a given multidimensional space. One of the main results of our research is, indeed, showing how these complex multidimensional big data analytics tools can be applied to real-life settings for effective and efficient healthcare analytics.

Indeed, the relevance of big data analytics applied to the healthcare setting has already emerged as one of the most exciting factors in recent big data research (e.g., [11]). Big data analytics in healthcare is a rapidly evolving field that leverages large and complex datasets to improve patient care, streamline operations, and advance medical research. This evident relevance is highlighted by many key points. First, big data analytics enables healthcare providers to analyze vast amounts of patient data in real-time, leading to *more accurate diagnoses and personalized treatment plans*. Second, healthcare organizations can use big data to *optimize their operations*, from managing patient flow to reducing wait times and improving resource allocation. This leads to cost savings and *better utilization of medical resources*. Third, big data analytics accelerates medical research by providing researchers with access to large datasets that can reveal patterns and correlations previously unnoticed, thus improving *medical research*. Lastly, by analyzing data from various sources, such as electronic health records and wearable devices, public health officials can monitor and respond to health trends and outbreaks more effectively. This is crucial for *managing epidemics and improving overall public health*.

This paper is an extension of the previous conference paper [12]. With respect to the previous contribution, in this paper, we significantly extend the experimental analysis of our analytical framework by also adding novel and meaningful case studies.

The remaining part of this paper is organized as follows. In Section 2, we provide an overview of our proposed architecture. Section 3 is devoted to the *PainRELifeDB* database, describing its anatomy and generative process. In Section 4, we describe the multidimensional modeling embedded in our Multidimensional Big Data Analytics methodology. Section 5 moves the attention to the design and implementation of the Multidimensional Big Data Analytics tools embedded in our architecture, along with their experimental evaluation against real-life datasets coming from the *Pain-RELife* research project. In Section 6, we provide our remarks about the outcomes of our research and future research directions. Finally, Section 7 reports conclusions and future work of our research.

2. A Cloud-Based Multidimensional Big Healthcare Data Analytics Framework: Anatomy and Main Layers

In this Section, we present the extensive reference architecture of the proposed framework for effective and efficient supporting Multidimensional Big Data Analytics; this clarifies the integration of specialized methods.

2.1. General Overview

We aim to provide a framework that can be used as a guide and help in extracting valuable insights and performing multidimensional data analysis in big healthcare settings.

Figure 1 illustrates our proposed architecture, which consists of three essential hierarchical levels, representing different steps of our approach to extract valuable and actionable information. It also shows data structures, models, and data information like companies, drugs, hospital structures, medical personnel, etc., as well as different tools that we use in the process.

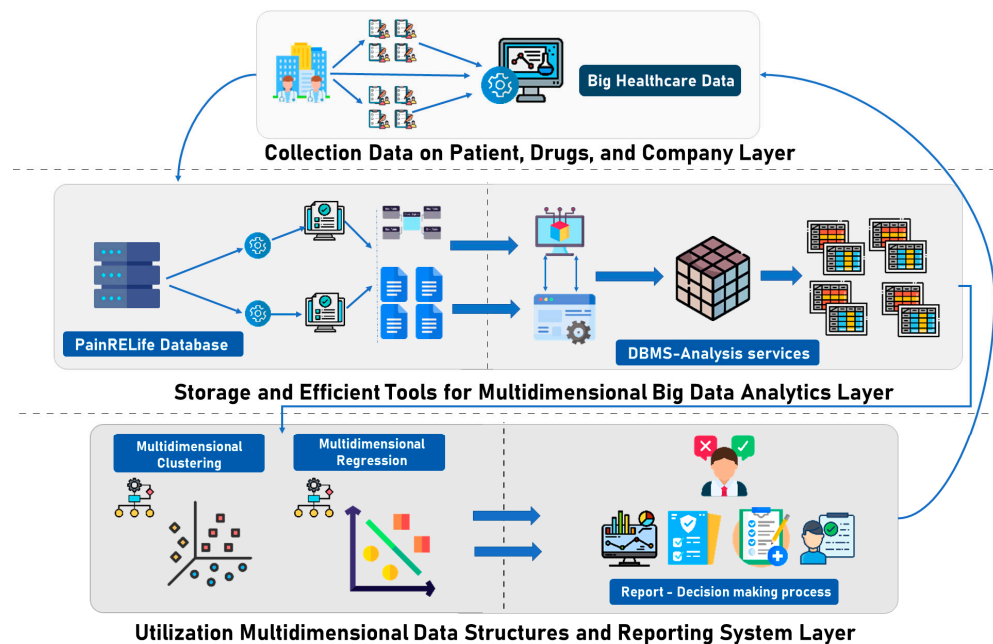


Figure 1. Cloud-Based Multidimensional Big Healthcare Data Analytics Architecture.

As shown in Figure 1, the proposed architecture demonstrates the usefulness of Multidimensional Big Data Analytics in specific use cases, such as healthcare. In summary, this real-life architecture consists of three fundamental layers that work together to enable an effective and efficient model for multidimensional big healthcare data analytics. In greater depth, the reference architecture is composed of three hierarchical tiers that incorporate several large-scale data/model components. Specifically, it is made of the following layers:

- **Big Data Collection Layer.** In this layer, collecting data is an essential step. Healthcare data is generated from different data sources with various types and formats. The focus is on obtaining data about specific individuals and patients, including their demographics, medical history, and other potential ailments. Additionally, data collection involves gathering information on other actors in healthcare domains, such as pharmaceutical products (e.g., drug composition, dosage, side effects, manufacturers, suppliers, and so forth). By capturing this data, users, providers, and companies involved in the target healthcare system can gain valuable insights, which help improve the healthcare sector (e.g., provide patients with specific treatments, improve supply chain management, etc.).
- **Big Data Storage Layer.** In this layer, all data from the previous layer are stored in *PainRELifeDB*, implemented on *Microsoft SQL Server*, which is used as a supporting data server for the analytical interpretation of multidimensional data, providing more realistic modeling for real-life databases that aim for high effectiveness, efficiency, and, notably, scalability across big data. In order to support multidimensional analytics, the so-called *Multidimensional Star Schema* [13], which enables us to perform *Online*

Analytical Processing (OLAP) [14] complex analysis operations such as slice-and-dice, drill-down, and roll-up on *multidimensional data cubes* (e.g., [13,15]), is introduced. Critically, all OLAP operations performed on these data cubes are subject to the security policies enforced by the *Security and Privacy capabilities* (see Section 2.2). For instance, *Access control rules* are translated into *dynamic query predicates* that filter multidimensional data based on the user's attributes and permissions, ensuring that sensitive *Protected Health Information* (PHI) is only accessible to authorized personnel with legitimate clinical or research needs (e.g., [16,17]). This prevents unauthorized data leakage and inference attacks during interactive analytical sessions. Therefore, deriving data cubes stores data in a structured format that is optimized for analytical queries, allowing for fast retrieval of aggregate data. Furthermore, *Microsoft SQL Server Analysis Services 2019* (SSAS) is an essential component in the Big Data Storage Layer. SSAS enables us to perform complex analysis on big data and gain valuable insights to support decision-making processes. Note that our framework applies a variety of optimization approaches, advanced tools, and technologies, including data partitioning, *Machine Learning* (ML), and *predictive modeling*. Additionally, *pivot analysis* (e.g., [18]) is performed as a starting point for the whole big data analytics process. This analysis provides an effective two-dimensional interface for exploring aggregate data in a multidimensional data model.

- *Multidimensional Reporting Layer*. The last layer is the front-end of our framework. It includes different types of analysis, which is mainly based on OLAP analysis integrated with the proposed Multidimensional Data Mining techniques, such as (i) Multidimensional Clustering, which is a crucial tool for analyzing large volumes of multidimensional data efficiently, and (ii) Multidimensional Regression, which is an ML method used for predictive modeling to discover new insights and predict trends by analyzing current data distributions. Furthermore, various analyses are performed based on discovered patterns and analytical results to gain valuable and actionable information (e.g., best drug, common side effects, and so forth). Different *data visualization tools* are used to deliver the results well graphically, which helps users understand complex patterns, trends, and outliers in the target healthcare repositories.

It is a matter of fact that, in sensitive sectors like healthcare, tools required to handle and interpret reporting data efficiently are mandatory for healthcare actors. Therefore, the proposed framework represents a meaningful step toward integrating advanced Multidimensional Big Data Analytics into clinical practice, paving the way for significant innovations in medical practices and providing a blueprint for future systems that aim to improve patient outcomes through data-driven insights.

Apart from analytics considerations, it should be noted that our proposed framework must also take into account *performance* aspects (e.g., [19]). Indeed, it has been proven that when big datasets are processed for analytics purposes, the relevant computational overheads are introduced, especially for the case of big multidimensional data (e.g., [20]). These drawbacks affect all the main big data management phases, including accessing, indexing, and querying, so that effective optimization solutions must be devised. As a consequence, specific *big data optimization design issues* (e.g., [21]) must be seriously taken into account. The latter is yet another contribution of our research.

2.2. Security and Privacy Aspects

Given the sensitive nature of PHI processed by our framework, dedicated *Security and Privacy capabilities* are integrated across all three architectural tiers, *Collection*, *Storage*, and *Reporting*, in order to ensure comprehensive protection against unauthorized access, data breaches, and cyber intrusions.

Among several *security and privacy mechanisms* (e.g., [22–26]), in our framework, we have opted to leverage the following ones:

- *Data Encryption*: All PHI stored in the *PainRELifEDB* database and its derived multi-dimensional data cubes are encrypted at rest using the *Advanced Encryption Standard* (AES-256) [26]. AES-256 is a *NIST-approved symmetric encryption algorithm* widely adopted in healthcare for its proven robustness and computational efficiency. Additionally, all data transmitted between the *Collection*, *Storage*, and *Reporting* layers are secured using *Transport Layer Security* (TLS) 1.3 protocols (e.g., [27]), in order to ensure protection against *spying* and *man-in-the-middle attacks* during OLAP query execution and data transfers.
- *Fine-Grained Access Control and Secure OLAP Query Processing*: To enforce the principle of least privilege, we implement *Attribute-Based Access Control* (ABAC) [22] within the *Big Data Storage Layer*. ABAC enables the definition of *dynamic, fine-grained access policies* based on user attributes. For example, a *general clinical practitioner* may only access aggregated pain statistics for their assigned patients, while a *hospital administrator* may view broader regional trends. This is complemented by *Role-Based Access Control* (RBAC) at the system level (e.g., [28]), managing user authentication and authorization for the SQL Server and SSAS instances. This multi-layered access control ensures that clinicians and analysts can only retrieve the specific multidimensional data subsets necessary for their clinical and administrative responsibilities, which effectively prevents *privilege abuse* and *unauthorized data exposure*.
- *Audit Logging and Blockchain Integration*: To ensure *non-repudiation* and *regulatory compliance*, all *Audit Logging and Blockchain Integration* data access events and OLAP query executions are recorded in a tamper-proof audit log. We integrate a *permissioned blockchain*, such as *Hyperledger Fabric* (e.g., [29]), to store these audit records immutably [7]. This blockchain-based logging provides an immutable and verifiable trail of all data interactions, which is essential for meeting the stringent audit requirements of regulations such as *General Data Protection Regulation* (GDPR) and *Health Insurance Portability and Accountability Act* (HIPAA) (e.g., [24,25]).
- *Privacy-Preserving Data Sharing*: In scenarios where data must be shared with external collaborators or research partners, our framework supports *privacy-preserving data sharing mechanisms* (e.g., [30–32]). Following the principles of the *Patient-Centric Blockchain Framework for Transparent and Secure Medical Data Sharing Using Modified AES* [26], we adopt a *patient-centric approach* where patients retain ownership and control over their data. *Modified AES encryption* [26] can be applied to shared datasets, ensuring that only authorized recipients with the correct decryption keys can access the information.

By embedding these multi-faceted security and privacy mechanisms (e.g., [22–26]) directly into the architecture, our framework transforms from a research prototype into a clinically viable platform capable of operating in real-life healthcare environments while maintaining the high analytical utility required for effective chronic pain management and research.

3. The *PainRELifEDB* Database

In this Section, we provide a detailed description of the design and the implementation of the database *PainRELifEDB*.

One partition of the target data is a *patient survey*, which is a questionnaire designed to evaluate patients' perceptions of pain. In more detail, patients are affected by two specific disease conditions: *stroke* and *early breast cancer*. All patients live in the same region, i.e., Lombardy. The goal of this survey is to collect valuable information about patients who suffer from *chronic pain*. The pain can be divided into two categories: (i) *Neuropathic*

pain: it is nerve pain that can happen if the nervous system malfunctions or is damaged; (ii) *Nociceptive pain*: it is a type of pain caused by damage to body tissue. In our data collection modeling, we extract and analyze patients' experiences reported before and after various medical treatments by providing questionnaires. The questionnaire data is stored in a separate table called *Survey*, which acts as a reference point for measures of the multidimensional modeling.

The sections within the table *Survey* consist of the following field types:

- *Pain Catastrophizing Scales* (PCS): The PCS scale consists of 13 attributes (PCS1–PCS13) that represent patient responses to their thoughts and feelings that may be connected with pain. The PCS is a widely validated psychological instrument used to assess catastrophic thinking related to pain experiences [33]. This attribute partition asks questioned patients to identify the *frequency* (on a scale from 0 to 4) with which they experience these thoughts and feelings while they are in pain.
- *Brief Pain Inventory* (BPI). BPI scale comprises 16 levels. It requires interviewing patients to indicate whether they have experienced pain. The BPI is a standardized questionnaire that evaluates both the intensity of pain (sensory dimension) and the degree to which pain interferes with daily functioning (reactive dimension) [34]. The aim is to evaluate the impact of the pain in relation to the performance of daily activities and emotional states.
- *Patients' Global Impression of Change* (PGIC). PGIC scale allows patients to describe their opinion of a probable change in pain (*Yes* or *No*), as well as a numerical value of this change (on a scale from 1 to 10). PGIC is a commonly used *single-item measure* that captures the patient's perceived overall improvement or deterioration following treatment [35].
- *Health Assessment* (SF). The SF scale requests that patients provide their perceptions of their emotional condition, performance of work activities, and health status. These metrics are derived from *Standardized Health Status Questionnaires* (e.g., SF-36) that assess physical functioning, role limitations, and general health perceptions [36].

We use Microsoft SQL Server as a primary tool to create and manage the *PainRELifedb* database. It allows users to perform various tasks such as creating, reading, and importing data, writing and executing queries, designing databases, managing security, and configuring server settings. Additionally, we use *SQL Server Management Studio* (SSMS) for the specific big data analytics tasks.

The *PainRELifedb* database is stored on a virtual infrastructure that runs on the VMware vSphere platform with a *three-physical-node cluster* (Foundation license). Each node is an ESXi host machine that features 256 GB RAM and two Intel(R) Xeon(R) Gold 5120 24 core CPU at 2.20 GHz with *Hyperthreading* (56 logical CPUs). In summary, over the three cluster nodes, 21 VMs are deployed. The three cluster nodes are designed for high reliability; hence, in case one of the three fails, all of the VMs in this node will be automatically restarted on one of the two remaining operational hosts.

In the following, we provide a detailed illustration of the database *PainRELifedb*, along with its dimensions and hierarchies. First, Figure 2 shows the percentage of disk space occupied by the database *PainRELifedb*, which corresponds to 2.72 GB, with the corresponding legend that shows how the space is allocated.

As shown by the *PainRELifedb* schema in Figure 3, several dimensions have been introduced in the *PainRELifedb* design to allow for a complete analysis of complex data, according to various perspectives of analysis. The database includes tables such as *Report*, *Physician*, *Patient*, *Pathology*, *Geography*, and *Drug*, each of which serves as a different OLAP *dimensional table* that contributes to multidimensional data analysis, along with the previous

table *Survey* that acts as a *fact table*. Note that we specifically designed the *PainRELifEDB* database to facilitate the integration of multidimensional data.

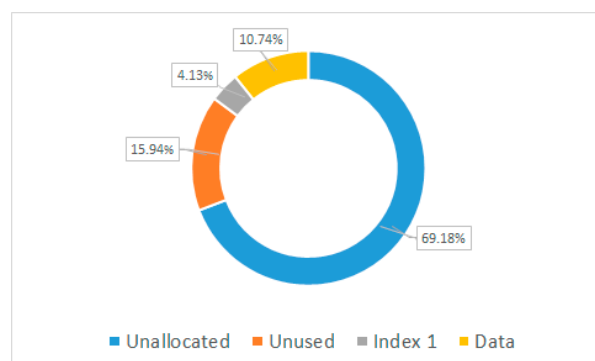


Figure 2. Space Occupancy for the Database *PainRELifEDB*.

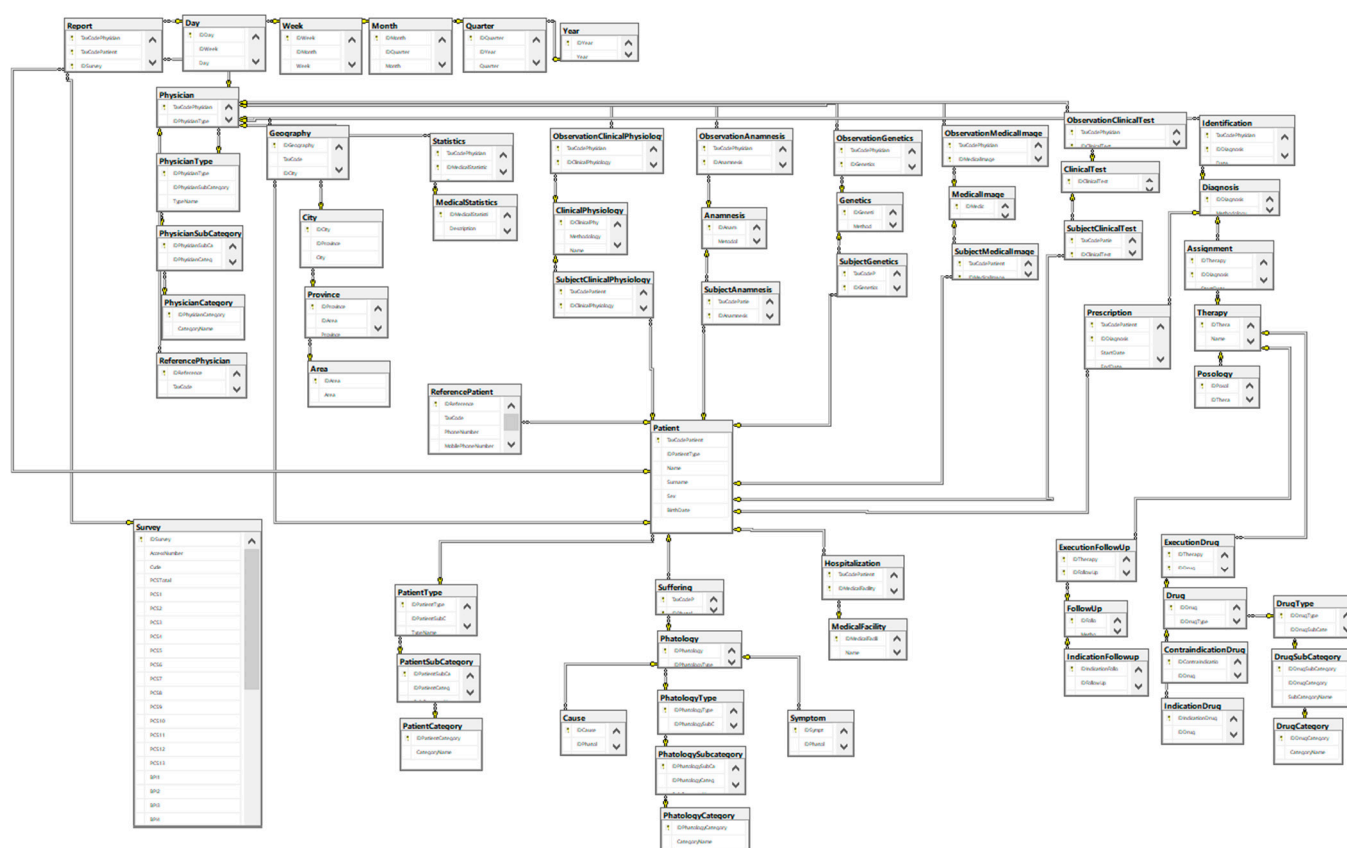


Figure 3. *PainRELifEDB* Schema.

Here, we report an example of a multidimensional dimension embedded in *PainRELifEDB*. The dimensional table *Report*, shown in Figure 4, contains crucial information about patients by storing data about the reports edited by patients over time. In this context, individual patients may be required to complete multiple questionnaires, which may affect data accessibility. Therefore, the date of the report must be recorded. In addition to this, to make it easier to organize and access this data, *Report* includes a hierarchical structure: *Day*, *Week*, *Month*, *Quarter*, and *Year*. Thanks to this hierarchical structure, it is possible to record, in fact, the particular day on which each survey is done, along with its OLAP dimensional hierarchy, which can then be modeled by hierarchical levels [37] not only for improved representation but also for enhanced multidimensional querying and analytics (e.g., [20]).

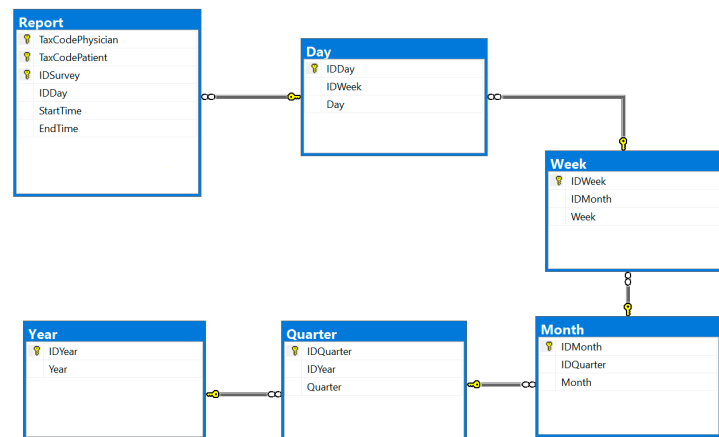


Figure 4. Dimensional Table Report.

The importance and usage of OLAP hierarchies in Multidimensional Big Data Analytics has been highlighted in several relevant research efforts (e.g., [14]). Indeed, in today's data-driven world, organizations are inundated with vast amounts of information from various sources. The challenge is not just to store this data but to transform it into actionable insights. This is where OLAP hierarchies become indispensable. In fact, on one side, OLAP hierarchies enable users to explore and analyze data in a structured and intuitive manner, thus ensuring a *structured exploration* of target data (e.g., [38]). Not by chance, this hierarchical structure makes it easier to identify patterns, trends, and anomalies that might be missed in a *flat* data model. On the other hand, OLAP hierarchies enhance the *performance of complex (OLAP-like) queries*, due to the fact that, by pre-aggregating data at various levels of the hierarchy, OLAP systems can quickly retrieve summarized information, reducing the need for time-consuming computations on raw data (e.g., [13]). This efficiency is crucial in big data environments where query response times can significantly impact decision-making processes (e.g., [39]). Lastly, OLAP hierarchies support multidimensional analysis by providing a framework to analyze data across different dimensions (e.g., [39]), thus fulfilling the requirement of examining data from multiple perspectives (i.e., *multiple dimensions*) of popular big data analytics systems.

In summary, the importance of OLAP hierarchies in Multidimensional Big Data Analytics cannot be overstated. They provide a robust foundation for exploring, analyzing, and visualizing complex (big) datasets.

4. Populating PainRELifeDB: Synthetic Data Generation

As mentioned in Section 1, PainRELifeDB database is populated with *synthetic data*, such that generated data relies on the integration of real-life data obtained from patient questionnaires and medical professionals coming from the real-life project Pain-RELife. In this Section, we focus on *synthetic data generation models and algorithms* introduced in our research.

4.1. Synthetic Data Generation Process

The adoption of synthetic data generation tools and algorithms is a relevant research innovation, and it is emerging more and more in the specific big data setting (e.g., [40]), even due to the obvious difficulties of acquiring *complete* real-life big datasets for non-commercial purposes (e.g., [41]). On the other hand, the quality, the accuracy, and the effectiveness of systems and tools developed on synthetic data (e.g., for simulation and tuning purposes) have already been largely proven (e.g., [42]).

Synthetic data are generated from real data using *algorithmic data generation techniques* via sampling, depending on the features and characteristics of the original (real) dataset (e.g., [43]). Moreover, there are *random approaches* that create random samples from probability distributions or random numbers within a predefined range. Furthermore, some methods use interpolation techniques to generate synthetic data to fill gaps in a dataset that contains missing values. Figure 5 shows the data generation process for *PainRELifedb*.

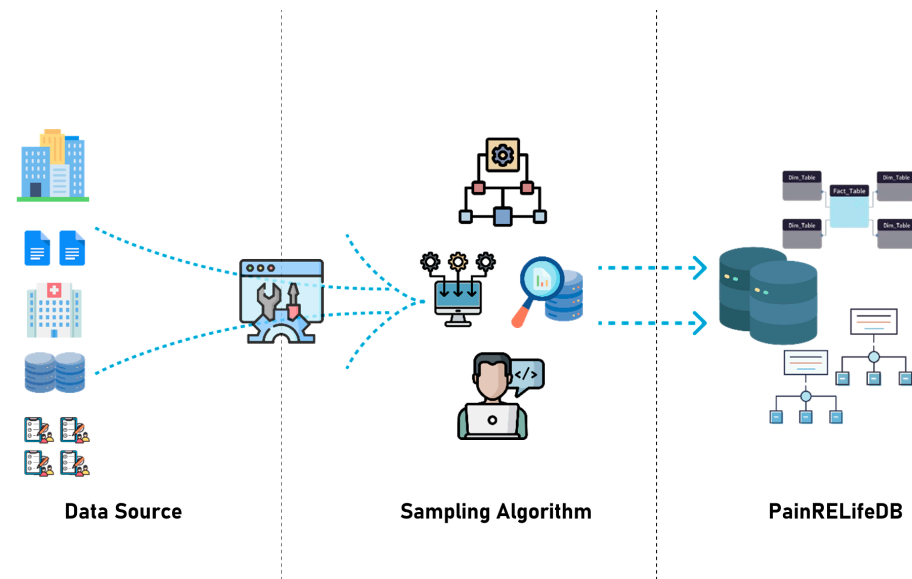


Figure 5. Algorithm for Populating the Database *PainRELifedb*.

Methods such as averaging, linear regression, and so forth are used to compute and estimate missing or incomplete values (e.g., [44]). Additionally, samples can be randomly generated from the real data and then manipulated to produce new data. ML models can also be used by taking the attributes of original data and producing new versions with the same style and distribution as the original data (e.g., [45]). Note that while producing synthetic data, it may not correctly reflect the complexity and diversity of the original data; therefore, dealing with the generated data requires much attention, and it is critical to analyze synthetic data and properly evaluate its quality before using it for a given goal.

Synthetic data may be used for different purposes, including data analysis, training ML algorithms, and testing new models. Generally, synthetic data can be used as a complementary or as a substitute for actual data. Specifically, when it is difficult or costly in time or money to gather real big data, generated data is beneficial in testing new models or algorithm capabilities, and exposes it to unique situations, scenarios, or events that do not regularly occur in reality or the data that rarely can appear in real life.

The first phase in our process is to collect patient surveys that include a range of information, such as patient demographics, medical history, and pain level. These questions are important in obtaining a thorough picture of the patient pain experience. The surveys are thoroughly examined to ensure that data are accurate and consistent. This stage is critical because it guarantees that data utilized to analyze the patient population are accurate and representative. Data are then organized and written in the *PainRELifedb* database, and various data fields from the surveys must be allocated to the respective columns in the database. Note that the *PainRELifedb* database was designed with scalability and data analysis flexibility in mind.

Here, we present our methods for producing synthetic data using real data as input and random sampling (e.g., [5]), which results in the *synthetic database* named *Synthetic PainRELifedb*. Figure 6 shows the steps and functions that we utilize to create data. In

particular, we utilize specific configurations based on the characteristics of each type of data to be generated. Essentially, our functions generate the final synthetic data using the same database, *PainRELifeDB*, which contains the real data from which the synthetic data are generated. The functions developed depend on the type of data to be created, namely *categorical data*, *numerical data* (which is further divided into *Integer data* and *Double data*), *temporal data*, and *Boolean data*. As shown in Figure 6, depending on the type of data, we generate random data via sampling and load the result into the final synthetic multidimensional database.

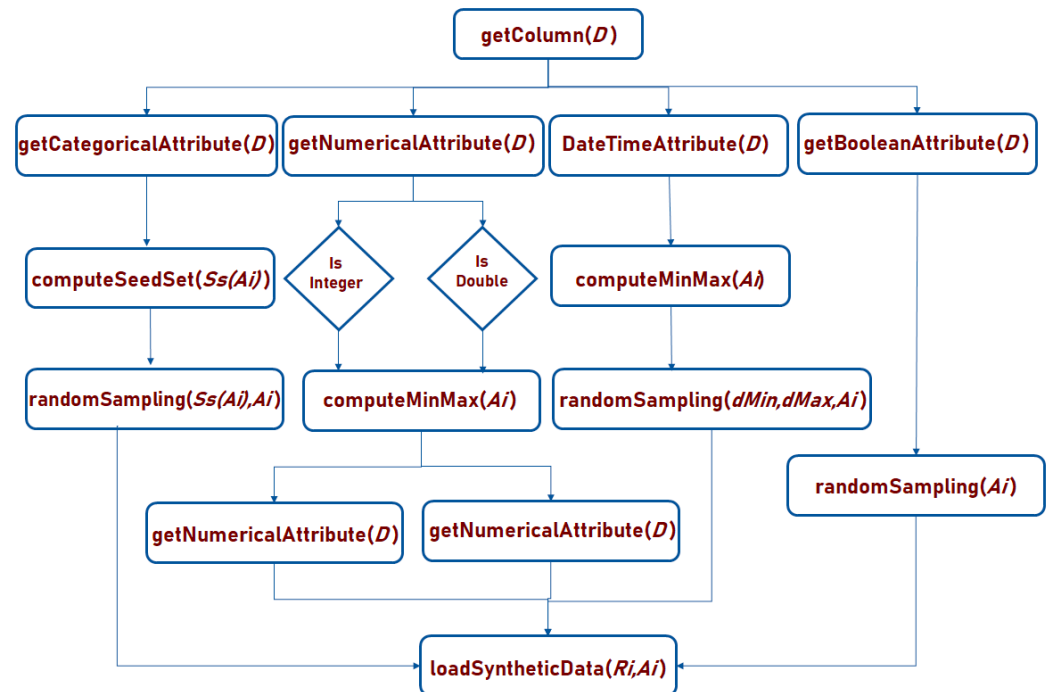


Figure 6. Generation of Synthetic Data for the Database *PainRELifeDB*.

The synthetic data generation process is driven by *advanced random sampling algorithms* (e.g., [5]) that operate by extracting sample distributions from the original real-life datasets and generating new instances adhering to these *multidimensional probability distributions*. To ensure the reliability of the generated data, a rigorous validation process is implemented (e.g., [46,47]). This process involves comparing the *statistical properties*, such as *mean*, *variance*, and *multivariate correlations*, of the synthetic dataset against the original real-life data. By evaluating these metrics, we guarantee that the synthetic data accurately preserves the original data patterns and clinical behaviors without introducing new, artificial, or biased patterns, thereby validating its effectiveness for downstream multidimensional analytics. More specifically, to ensure the reliability of the analytical results, the synthetic data generation process in *PainRELifeDB* was not purely stochastic. We employed a *stratified random sampling* approach (e.g., [48,49]) based on the *Probability Distribution Functions* observed in the preliminary real-life pilot data from the Lombardy region. Specifically, key correlations between patient demographics (age, gender) and pain intensity (*Numerical Rating Scale—NRS*) were preserved by applying *conditional probability constraints* during the generation phase, ensuring that the synthetic population maintains the statistical essence of the clinical reality while fully respecting patient privacy and GDPR requirements.

4.2. Validation of Synthetic *PainRELifeDB*

To substantiate the fidelity of our synthetic data generation process, we performed a comprehensive quantitative validation comparing *Synthetic PainRELifeDB* against the origi-

nal real-life dataset from the *Lombardy* region. This validation ensures that the analytical results derived from the synthetic data are generalizable to genuine clinical populations and are not biased by *Artificial Generation Artifacts*.

For each continuous clinical variable, including *Age*, *NRS*, *BPI*, and *PCS scores*, we applied the *two-sample Kolmogorov–Smirnov* (KS) test (e.g., [50,51]). The KS test measures the maximum absolute difference between *Empirical Cumulative Distribution Functions* (ECDFs). Given two independent samples (real and synthetic) of sizes n_1 and n_2 , the two-sample KS statistic is defined as follows:

$$D_{n_1, n_2} = \max |F_{real} - F_{synth}| \quad (1)$$

where F_{real} and F_{synth} are the empirical cumulative distribution functions of the real and synthetic samples, respectively. The null hypothesis posits that the synthetic and real samples are drawn from the same continuous distribution. The test statistic, denoted as T , is then defined as follows:

$$T = D_{n_1, n_2} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}} \quad (2)$$

The corresponding *p-Value* is computed using the *Kolmogorov distribution*, which approximates the probability of observing a test statistic at least as extreme as T under the null hypothesis (e.g., [50,51]).

Across all tested variables, the resulting *p-values* consistently exceeded the 0.05 significance threshold, which confirms that the synthetic distributions are statistically indistinguishable from their real-life counterparts. Table 1 reports the KS test results for the most clinically relevant variables.

Table 1. Kolmogorov–Smirnov Test Results for Key Clinical Variables in *PainRELifEDB*.

Variable	KS Statistics (D)	p -Value
<i>Patient Age</i>	0.084	0.73
<i>BPI Average Pain</i>	0.071	0.89
<i>BPI Interference</i>	0.092	0.62
<i>PCS Total Score</i>	0.078	0.81
<i>PCS Rumination</i>	0.065	0.94
<i>PCS Magnification</i>	0.103	0.47
<i>PCS Helplessness</i>	0.087	0.69
<i>NRS Pain Intensity</i>	0.095	0.58

It should be noted that the high fidelity achieved is a direct result of our *stratified random sampling methodology* (e.g., [48,49]), which was constrained by the empirical *Probability Distribution Functions* observed in the original data. Critically, we applied *conditional probability constraints*, for example, ensuring that the joint distribution of age, gender, and pain intensity in the synthetic population mirrors the clinical reality observed in the Lombardy cohort. This approach prevents the introduction of implausible patient profiles while fully preserving the multivariate essence required for reliable healthcare analytics.

By providing this multi-faceted quantitative validation, we confirm that the synthetic *PainRELifEDB* is a robust proxy for real clinical data.

5. Design and Implementation of Multidimensional Big Data Analytics Tools

In this Section, we focus attention on the design and implementation of Multidimensional Big Data Analytics tools embedded in our proposed architecture.

5.1. Experimental Setup

To ensure the reproducibility and transparency of our experimental evaluation, we provide here a comprehensive description of the data scale, computational infrastructure, and algorithm-specific parameters employed throughout our research.

The synthetic *PainRELifeDB* database comprises approximately 2.72 GB of data (see Figure 2), with the central fact table (Survey) containing over 1.2 million records. This scale is representative of typical real-life healthcare repositories at the regional hospital network level, ensuring that our performance and scalability results are practically relevant.

Regarding the computational infrastructure, the *PainRELifeDB* database and all analytical computations were hosted on a *VMware vSphere* virtualized cluster consisting of three physical *ESXi host nodes*. Each node features 256 GB RAM and two Intel(R) Xeon(R) Gold 5120 24-core CPUs at 2.20 GHz with *Hyperthreading* (totaling 56 logical CPUs per node). Across the three cluster nodes, 21 *Virtual Machines* (VMs) are deployed, with dedicated resource allocation for the Microsoft SQL Server instance hosting *PainRELifeDB*. The cluster is configured for high availability, such that if any node fails, its VMs are automatically restarted on the remaining operational hosts.

5.2. Experimental Results

Figure 7 shows the conceptual *Dimensional Fact Model* (DFM) [52] through which it is possible to represent data within the main OLAP data cube over *PainRELifeDB*.

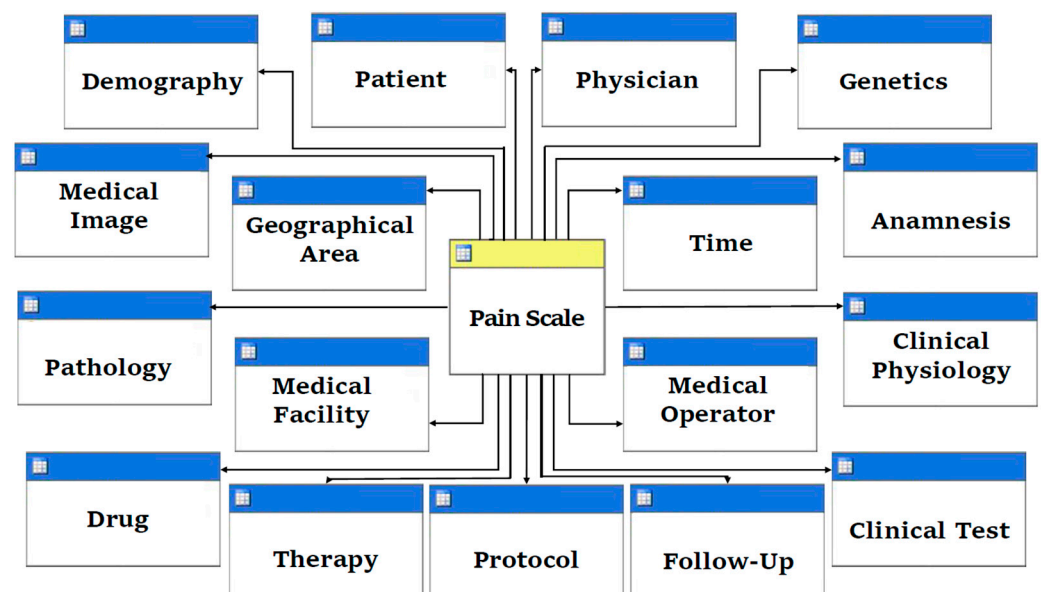


Figure 7. DFM over the Database *PainRELifeDB*.

This model captures the dimensions of analysis and the relevant facts pertaining to the modeled reality. The multidimensional model offers several key advantages: (i) provides support for conceptual design; (ii) creates an environment for intuitive and guided querying. In terms of graphical representation, the DFM is depicted as a star schema. At the center of the diagram, *Pain Scale*, also referred to as *Survey*, holds prominence in the decision-making process. The measure *Survey* is intricately linked to 17 dimensions by establishing relationships that reflect the analytical dimensions of the model. The multidimensional data analysis involves utilizing prominent analysis and reporting tools, such as *Microsoft Excel*. These tools aim to simplify the analysis process by providing specific functions and formulas and enabling the creation of relevant visualizations. In addition, the analysis encompasses the use of two specific Data Mining structures. The first structure, known

as Multidimensional Clustering, clusters similar elements into distinct groups or clusters based on shared characteristics and attribute values in a multidimensional space.

This clustering process facilitates the differentiation of clusters based on their proximity or distance from one another. The second structure, Multidimensional Regression, is used in the application domain of ML, an important branch of artificial intelligence. ML focuses on a machine's ability to generalize experiences and perform inductive reasoning. Notably, Multidimensional Regression demonstrates its strength in autonomously generating examples, which it achieves by training on a set of learning data. By employing the DFM and utilizing the aforementioned structures, we can benefit from a comprehensive and efficient approach to data analysis.

The selection of the underlying algorithms for both *Multidimensional Clustering* and *Multidimensional Regression* is carefully driven by the specific characteristics of multidimensional healthcare data, particularly the high-dimensionality and heterogeneous measurement scales (e.g., [53]).

For the Multidimensional Clustering task, we adopted the *K-Means++ algorithm* [54]. This choice is motivated by its computational efficiency and scalability; its time complexity scales linearly with both the *number of data points* and the *number of dimensions*, making it well-suited for large healthcare data cubes. Critically, prior to clustering, we apply *normalization* (e.g., [55]) to all continuous dimensions (e.g., *Pain Scores*, *Age*, and *PCS/BPI scales*). This normalization step is indispensable in multidimensional settings as it ensures that all dimensions contribute equally to clustering metrics. Without this normalization, variables with larger numerical ranges (e.g., *Age*) would dominate the distance computations. To determine the *optimal number of clusters*, k , we combined the *Elbow Method* (e.g., [56]) with the *Silhouette Coefficient* (e.g., [57]). The Silhouette Coefficient, which quantifies both *intra-cluster cohesion* and *inter-cluster separation*, serves as a critical validation metric in high-dimensional spaces.

Moreover, for the Multidimensional Regression task, we implemented a *Multivariate Linear Regression model* (e.g., [58]) with *Lasso (L1) regularization* [59]. In the context of multidimensional data, multivariate regression enables the simultaneous prediction of multiple correlated outcome measures (e.g., different BPI pain dimensions). The addition of L1 regularization is particularly beneficial for healthcare analytics, as it performs *automatic feature selection* by driving the coefficients of irrelevant or redundant predictors to zero. Given that our multidimensional data cubes contain numerous highly correlated clinical variables, Lasso regularization effectively identifies the most influential predictors of pain progression while preventing overfitting and preserving model interpretability. Model performance was evaluated using standard metrics, including *Root Mean Square Error* (RMSE) and *R-Squared* (R^2) metrics (e.g., [60,61]), which confirmed the model's strong predictive power and reliability across the multidimensional feature space.

According to a more general interpretation, the primary aim of multidimensional analysis is to extract pertinent information from a substantial volume of data, which, taken individually, could not be considered suitable for satisfying more in-depth analyses. However, analyzing the main data cube in its entirety could cause anomalies that would disfavor or, in the worst-case scenario, prevent a very specific manipulation of the data itself.

Consequently, the main data cube is partitioned into three distinct target cubes, each representing specific dimensions. In our experiments, we provide results on the following data cubes: (i) *PatientCube*; (ii) *PhysicianCube*; (iii) *PathologyCube*. It should be noted here that these data cubes are effective and efficient in assessing the capabilities of our proposed multidimensional big data analytics framework since they touch various perspectives of analysis one can find in real-life healthcare settings. Not by chance, the first data cube focuses on patients, the second on physicians, and the third on pathologies. The latter are

fundamental perspectives of the whole analytical domain, on which further analysis may also be devised.

First, *PatientCube*, shown in Figure 8, is designed to store and organize data specifically related to patients. The fact table *Survey* contains various key attributes and measures related to patient information. *Survey* is linked to important dimensions such as *Patient*, *Geography*, *Anamnesis*, *Diagnosis*, and *Report*. *Patient* is further divided into several hierarchies to provide more detailed analysis of patient data. The hierarchies include *PatientType*, *PatientSubCategory*, and *PatientCategory*, allowing for different levels of classification and segmentation of patients. *Report* focuses on the temporal aspect of patient data. Like *Patient*, *Report* features a hierarchy that includes *Day*, *Week*, *Month*, *Quarter*, and *Year*. This enables analysis of patient information based on different time intervals. *Geography*, on the other hand, allows for analysis of patient data based on geographic locations. It includes a hierarchy that regards *City*, *Province*, and *Area*, providing insight into the distribution of patients in different geographic regions.

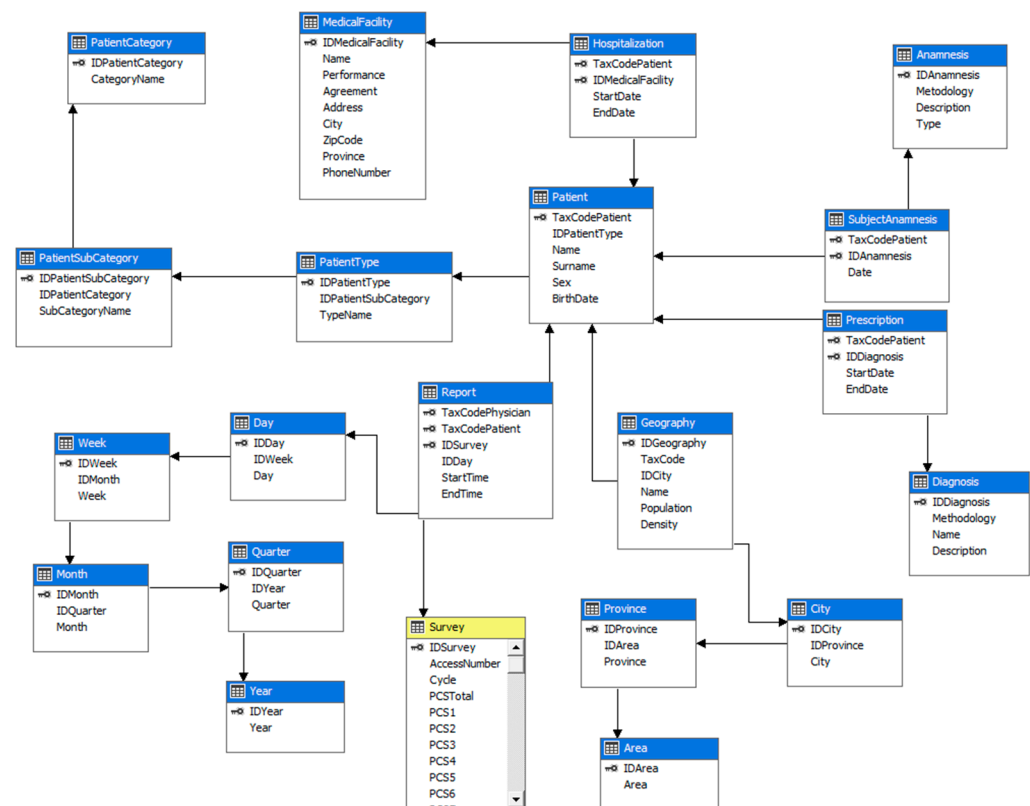


Figure 8. Data Cube *PatientCube*.

Through *PatientCube*, we can extract specific patient-related information from *PainRE-LifeDB*. This ensures that only data related to patients is included and that information about doctors or other entities are excluded. Any analysis or relationships related to doctors will be addressed separately in another data cube.

We aim to gather crucial information about patients through the *PatientCube*. A comprehensive evaluation of medical well-being cannot disregard the subject's medical history, the facility in which they were hospitalized, or the diagnoses provided by health-care professionals. Additionally, for statistical and informative purposes, it is valuable to gather geographical details concerning patients by including their city of origin, province, and reference area. The analysis will delve deep into extracting data regarding patients' specific concerns, as well as the impact that diagnosed conditions have had on various aspects of their lives.

We have performed a performance optimization on *PatientCube*, where we pre-computed aggregations of the data cube in order to provide *faster query response times*, since query execution is the baseline component of higher big data analytics procedures (e.g., [62]); thus, achieving optimization in this phase will result in a consequent improvement of spatial and temporal complexity of those procedures. To this end, in our study, we vary the *Number of Aggregations* of the target data cube, and we retrieve the corresponding *Performance Gain* in terms of the *average query response time* of a *random population of aggregate queries* over the *optimized data cube* with respect to the same metrics over the original data cube.

Our goal is to find the *optimal* number of aggregations. This number not only provides satisfactory response time but also prevents excessive partition size. A greater number of aggregations produces faster response times, but it also requires more storage space and may take longer to compute. Moreover, as we compute more aggregations, earlier aggregations produce considerably larger performance gains than later aggregations. The main benefit is that the reduction of less useful aggregations increases performance as well, along with the ability to control the number of aggregations to be computed. In addition, as mentioned, the final big data analytics processes benefit from the nice performance gain.

Following that, we test and stress a set of synthetic random range queries Q_S against the original data cube *PatientCube* with a specific number of aggregations, such that each query $q_i \in Q_S$ is defined on a given multidimensional range R and for a given aggregation operator A , as follows:

$$Q = \langle R, A \rangle \quad (3)$$

The average query execution time T_{AVG} due to the execution of queries in Q_S over the original data cube, *PatientCube*, is thus defined as follows:

$$T_{AVG} = \frac{1}{\|Q_S\|} \sum_{i=0}^{\|Q_S\|-1} T(q_i) \quad (4)$$

where $T(q_i)$ is the query execution time of $q_i \in Q_S$ and $\|Q_S\|$ is the *arity* of Q_S .

Then, by computing the *same* metrics over the optimized data cube *PatientCube*, given a number of aggregations $n_{agg,i}$, we obtain the average query execution time $T_{AVG}(n_{agg,i})$ for the *same* population of random queries, namely $T_{AVG}(n_{agg,i})$, and we finally compute the performance gain $T_{PG}(n_{agg,i})$ due to $n_{agg,i}$ as follows:

$$T_{PG}(n_{agg,i}) = \frac{|T_{AVG}(n_{agg,i}) - T_{AVG}|}{T_{AVG}} \quad (5)$$

Figure 9 shows the performance gain with respect to the number of aggregations on *PatientCube*. Indeed, in this optimization experiment, the *Number of Aggregations* refers to the count of *pre-computed aggregation cells* generated for each data cube (e.g., [63]). These aggregations correspond to different combinations of dimensional hierarchy levels (e.g., aggregating by *Month* instead of *Day*, or by *Province* instead of *City*). The number is a *unitless count* representing the size of the pre-computed aggregation space. In our case, we evaluated configurations ranging from 500 to 4000 aggregations to identify the optimal trade-off between *query performance gain* and *storage/computation overhead*.

As shown in Figure 9, we obtain up to 21% performance gain in query response time for a number of aggregations equal to 4000. Derived results demonstrate that an effective gain can be obtained when optimizing the number of aggregations of the target data cubes, yet taking into consideration space/time tradeoffs for big data processing (e.g., [64]).

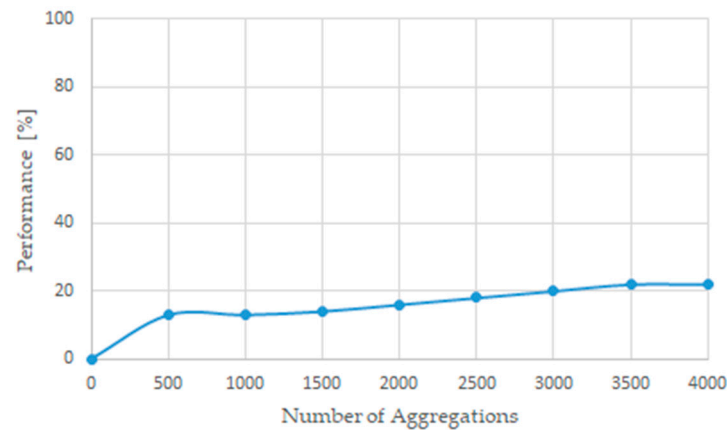


Figure 9. Aggregation Performance Gain on *PatientCube*.

From a *practical clinical and analytical perspective*, this 21% performance gain translates into enhanced responsiveness for real-time *Clinical Decision Support Systems* (CDSS). When clinicians manipulate large-scale multidimensional healthcare repositories during active patient consultations, such as executing interactive *Drill-Down* or *Slice-and-Dice* operations across complex pain metrics (e.g., BPI, PCS scales), any noticeable computational latency can disrupt clinical workflows. By reducing query execution times, our framework enables interactive, sub-second analytical feedback. This allows medical professionals to smoothly explore complex multi-attribute patient profiles, rapidly formulate diagnostic hypotheses, and personalize chronic pain management plans without computational bottlenecks impeding the *point-of-care experience*.

Without loss of generality, our aim is to gather valuable insights about patients by analyzing the data present in the data cube *PatientCube*. By examining the dimensions of *Patient* and *Geography*, specifically focusing on the attribute *City* within the hierarchy *Geography*, we can determine which cities have the highest and lowest number of patients. Figure 10 shows the histogram that reports the results. According to Figure 10, *Brescia* stands out as the city of origin for the largest number of patients, with a count of 117. Conversely, both *Lodi* and *Sondrio* have an equal number of patients, which happens to be the lowest in comparison, with 80 patients each.

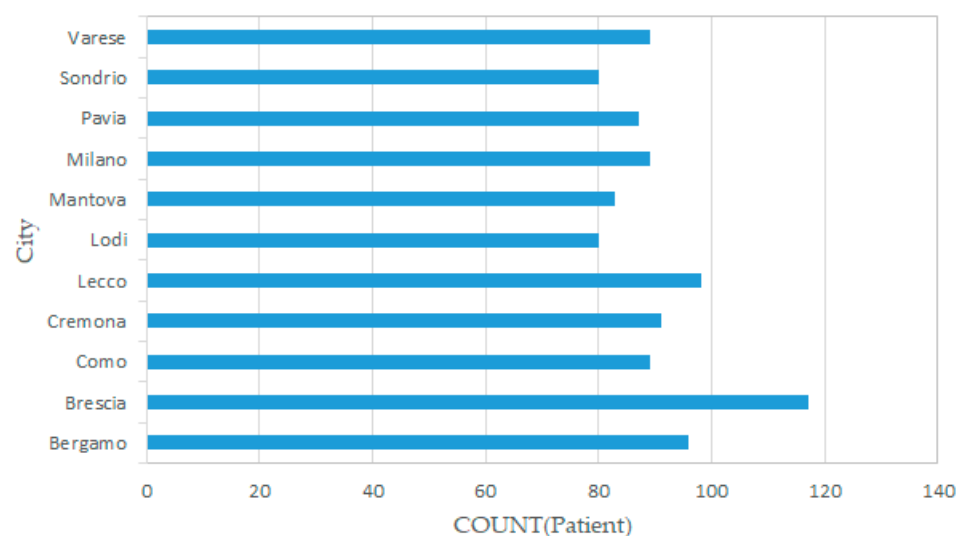


Figure 10. *PatientCube*_MDPivoting on *PatientCube*.

From Figure 10, it clearly emerges that the *power* of analytics (e.g., [65]) is especially evident in the emerging big data context. Indeed, thanks to aggregates, it is possible to

discover *trendy patterns* in target big datasets, which, on the contrary, would not be possible to achieve if other analytical, e.g., item-based, methodologies were adopted to this end. This evident phenomenon is “naturally” magnified when big data are considered, due to the well-known characteristics of these kinds of data, for which *big data understanding*, *exploration*, and *fruition* become a relevant research challenge (e.g., [38]).

Multidimensional Clustering analysis is carried out as well. In Figure 11, the clustering is made with respect to the variable *Population* of *PatientCube*. *Cluster 1* is the largest in terms of population, equal to 108. Followed, in descending order, by *Cluster 3* with 84, *Cluster 2* with 82, *Cluster 5* with 80, *Cluster 4* with 76, *Cluster 6* with 67, *Cluster 8* with 61, *Cluster 7* with 56, *Cluster 9* with 51, and *Cluster 10* with 35.

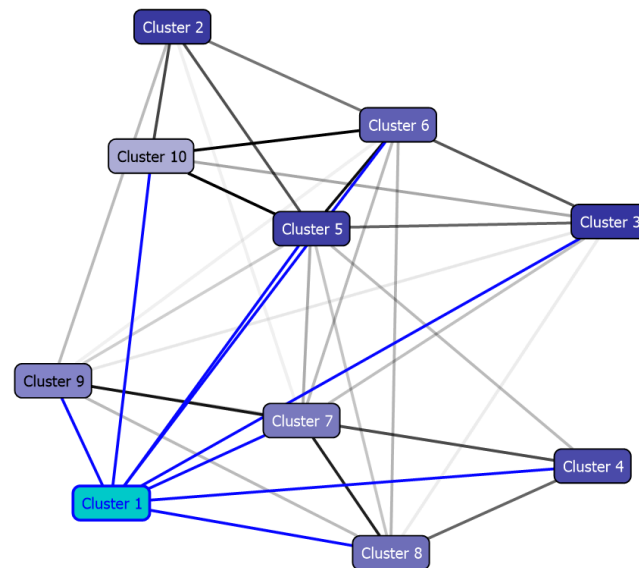


Figure 11. *PatientCube_MDClustering on PatientCube—Cluster Variable: Population.*

Figure 12 shows the *cluster profiles* for the clustering of Figure 11, which summarizes in a simplistic and graphically intuitive way data relating to the various variables that the previous cluster diagrams shown have indicated. Figure 13 shows the *cluster discrimination* analysis in which values present in two different clusters are compared. In particular, in Figure 13, *Cluster 1* and *Cluster 2* are compared, as the differences shown are common to all others. Relevant differences, therefore, concern female patients affected by *breast cancer* in *Cluster 2* and male patients affected by *stroke* in *Cluster 1*. The severity status is *intermediate* as regards *Cluster 1*, while not at all in *Cluster 2*. As regards *hospitalization* rate, it is present at about 15% in favor of *Cluster 1* compared to 15% of the *domiciliation* rate in *Cluster 2*.

It should be noted here that Clustering is a powerful methodology for analyzing real-life datasets. On the other hand, when combined with multidimensional paradigms (e.g., [66]), it further enhances its expressive power and its accuracy in advanced analytics methodologies.

Beyond *statistical cluster descriptions*, here, we provide a clinical interpretation of the top three most clinically meaningful clusters identified in the *PatientCube* analysis (e.g., [67–69]), which are the following ones:

- *Cluster 1* (*Population* = 108). This cluster is predominantly characterized by patients suffering from *stroke-related chronic pain*, with a higher proportion of males (62%) and an *intermediate severity status*. Patients in this cluster report moderate BPI scores (average BPI interference = 5.7/10) and relatively low PCS scores (average = 18.4/30), which suggests lower levels of pain catastrophizing compared to other clusters. Notably, this cluster exhibits a 15% *hospitalization* rate, indicating that these patients are more likely

to require inpatient care. From a clinical management perspective, this cluster may benefit from integrated rehabilitation programs that address both motor and sensory deficits, with a focus on functional restoration rather than purely pharmacological pain relief. The predominance of male patients with stroke suggests that *gender-specific rehabilitation strategies* may be warranted.

- **Cluster 2 (Population = 82).** This cluster is predominantly characterized by patients suffering from *breast cancer-related chronic pain*, with a higher proportion of females (approximately 78%) and a *severe pain status*. Patients in this cluster report significantly higher BPI scores (average BPI interference = 8.2/10) and elevated PCS scores (average = 26.7/30), indicating high levels of pain catastrophizing. Importantly, this cluster exhibits a 15% *domiciliation rate*, suggesting that these patients are more likely to receive care at home rather than in hospital settings. This pattern may reflect the growing trend toward ambulatory and home-based cancer care. From a clinical perspective, this cluster may benefit from *multimodal pain management approaches* that combine pharmacological interventions (e.g., opioids, adjuvant analgesics) with psychological support (e.g., cognitive-behavioral therapy) to address the high catastrophizing levels. The predominance of female patients with breast cancer also highlights the need for *gender-sensitive pain management protocols* that consider the psychosocial dimensions of cancer-related pain.
- **Cluster 3 (Population = 84).** This cluster is characterized by patients with *mixed pathologies* (both stroke and breast cancer) and a *diverse severity profile*. Notably, this cluster exhibits the *highest geographical diversity*, with patients distributed across all cities in the Lombardy region. This suggests that the clustering may be influenced by regional differences in healthcare access, referral patterns, or treatment availability. From a clinical management perspective, this cluster represents the most heterogeneous patient group, requiring personalized care plans that consider both the underlying pathology and the local healthcare resources available. The diversity of this cluster underscores the importance of *flexible, adaptable care pathways* that can be tailored to individual patient needs and regional contexts.

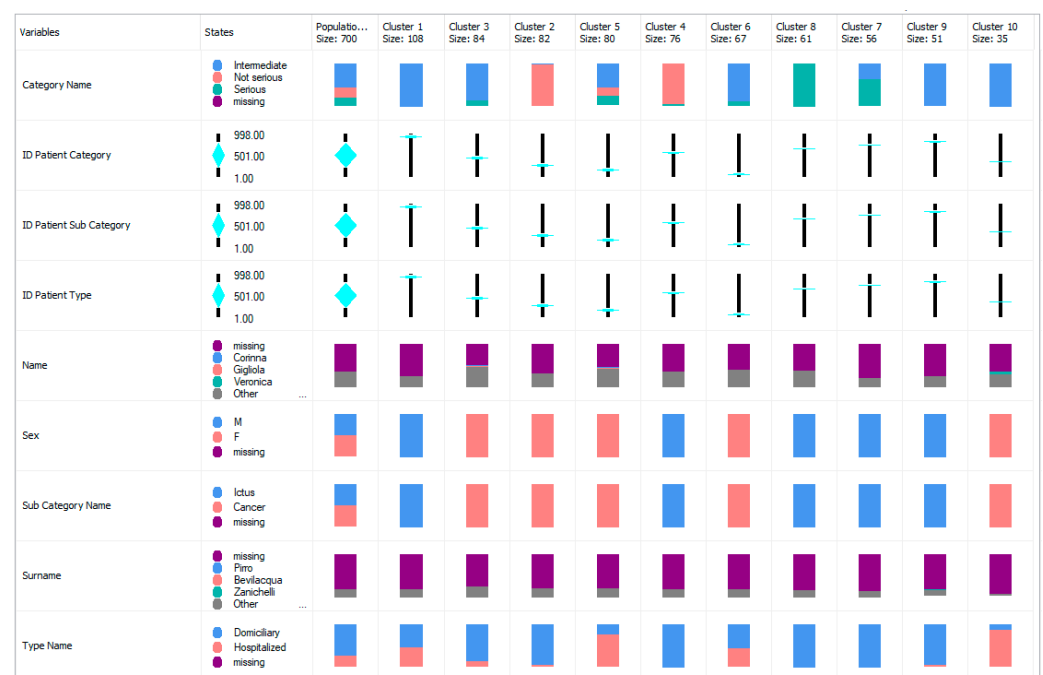


Figure 12. PatientCube_MDClustering on PatientCube—Cluster Profiles.

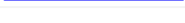
Variables	Values	Favors Cluster 1	Favors Cluster 2
ID Patient Type	1 - 552		
ID Patient Sub Category	553 - 998		
ID Patient Sub Category	1 - 552		
Sex	F		
Sex	M		
Sub Category Name	Cancer		
ID Patient Type	553 - 998		
Sub Category Name	Ictus		
ID Patient Category	553 - 998		
ID Patient Category	1 - 552		
Category Name	Intermediate		
Category Name	Not serious		
Type Name	Hospitalized		
Type Name	Domiciliary		
Birth Date	10/12/1971 3:39:38 PM - 5/21/1973...		
Birth Date	5/21/1973 6:35:30 PM - 12/26/1978...		

Figure 13. *PatientCube_MDClustering on PatientCube—Cluster Discrimination between Cluster 1 and Cluster 2.*

Indeed, the cluster profiles reveal three distinct patient subgroups with different pain characteristics, psychological profiles, and care utilization patterns. Cluster 1 (stroke, male, moderate pain, higher hospitalization) suggests a need for integrated rehabilitation programs and functional restoration. Cluster 2 (breast cancer, female, severe pain, high catastrophizing, domiciliation) suggests a need for multimodal approaches combining pharmacological and psychological interventions. Cluster 3 (mixed pathologies, geographically diverse) suggests a need for personalized, regionally adapted care pathways. These findings provide actionable insights for healthcare providers and policymakers seeking to optimize resource allocation and personalize treatment strategies for patients with chronic pain.

Next, we perform the Multidimensional Regression analysis over *PatientCube*. *Mining Multidimensional Regression Framework* is able to predict what could be future values of the measures of pain based on the trend of historical values present in *PainRELifedb*. Figure 14 shows measures to which a different color is associated.

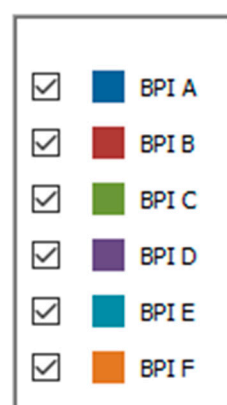


Figure 14. *PatientCube_MDRegression on PatientCube—Measures.*

In the context of the predictive analysis performed via Multidimensional Regression, it is important to clarify the meaning of the reported numerical values and percentages. The reported percentages represent the relative rate of change (either an increase or a decrease) of the specific target measure compared to its historical baseline within the multidimensional space. To rigorously evaluate the performance and reliability of these predictive models, *standard prediction accuracy metrics* (e.g., [70]), such as the *Root Mean*

Square Error (RMSE) and *Mean Absolute Error* (MAE), were evaluated. These performance metrics confirm the high accuracy and robust predictive power of our Multidimensional Regression tool when applied to complex healthcare data distributions, ensuring that the predicted trends closely align with expected clinical trajectories.

Figure 15 shows predictive analysis based on the previously selected measures. Analysis reports percentages of the last current values, represented by the last non-dashed values, and percentages of the last predicted values. The percentages are the following:

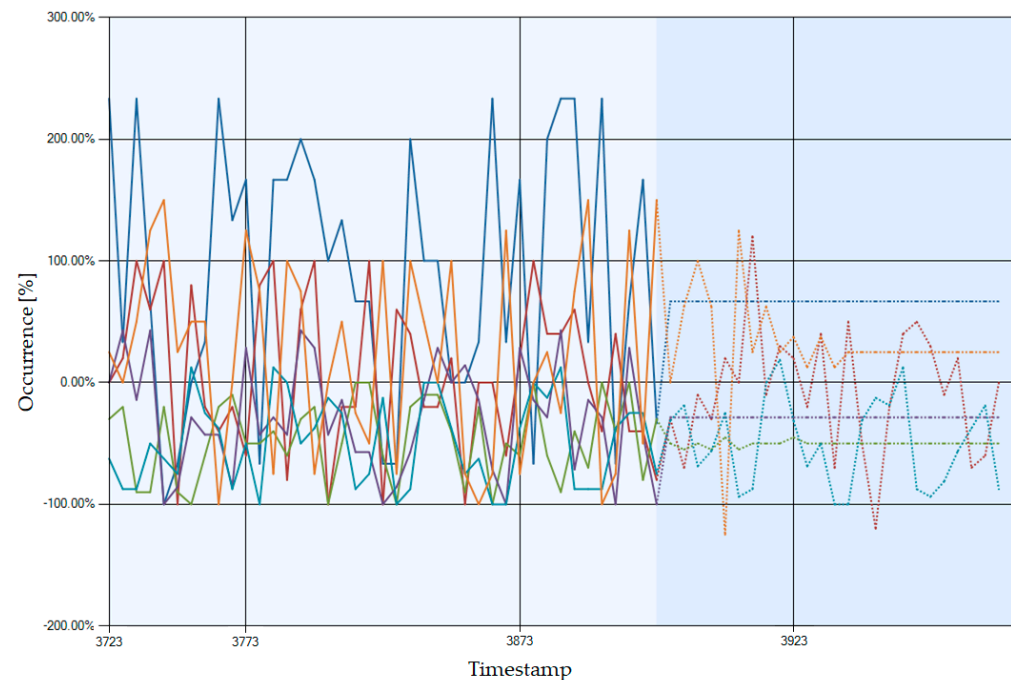


Figure 15. *PatientCube_MDRRegression on PatientCube* based on the Measure reported in Figure 14.

- BPI A (blue): current value −34%, last predicted value 67%;
- BPI B (red): current value −80%, last predicted value 0%;
- BPI C (green): current value −30%, last predicted value −50%;
- BPI D (purple): current value −100%, last predicted value −28%;
- BPI E (light blue): current value −75%, last predicted value −87%;
- BPI F (orange): current value −29%, last predicted value −62%.

Secondly, *PhysicianCube*, as shown in Figure 16, is designed to store and organize data about physicians' patient medical reports and statistical records. The core structure of the data cube *PhysicianCube* involves a central fact table named *Survey*, which is intricately linked to various important dimensions such as *Physician*, *Statistics*, *Hospitalization*, and *Report*. These dimensions facilitate a comprehensive representation of data, making it possible to conduct detailed analysis from different perspectives. To enhance the depth of patient data analysis, the dimension *Patient* is further subdivided into several hierarchical levels, each offering different levels of classification and segmentation.

These hierarchies include *PhysicianType*, *PhysicianSubCategory*, and *PhysicianCategory*, which together allow for more nuanced categorization of physicians based on specific criteria. The so-defined hierarchical structure enables the system to efficiently process and segment physicians according to their roles and specialties.

In a similar manner to the case of *PatientCube* (see Figure 9), Figure 17 shows the result of the query response time analysis for better data storage optimization, on the basis of the number of aggregations on *PhysicianCube*. Here, it is possible to notice that a performance gain in query response time of 30% can be achieved when a corresponding

number of aggregations is equal to 3500 or, very closely, 4000. The dimension *Statistics* primarily focuses on the statistical aspects of patient data cared for by physicians, with a particular emphasis on factors such as *vitality* and *pathological* information. Similar to the dimension *MedicalStatistics*, which extends the *Physician* hierarchy, the dimension *Hospitalization*, which is part of the hierarchy of the dimension *Patient*, spans down to *MedicalFacility*. This extended hierarchy provides the ability to analyze hospitalization data and track patient care across various medical facilities.

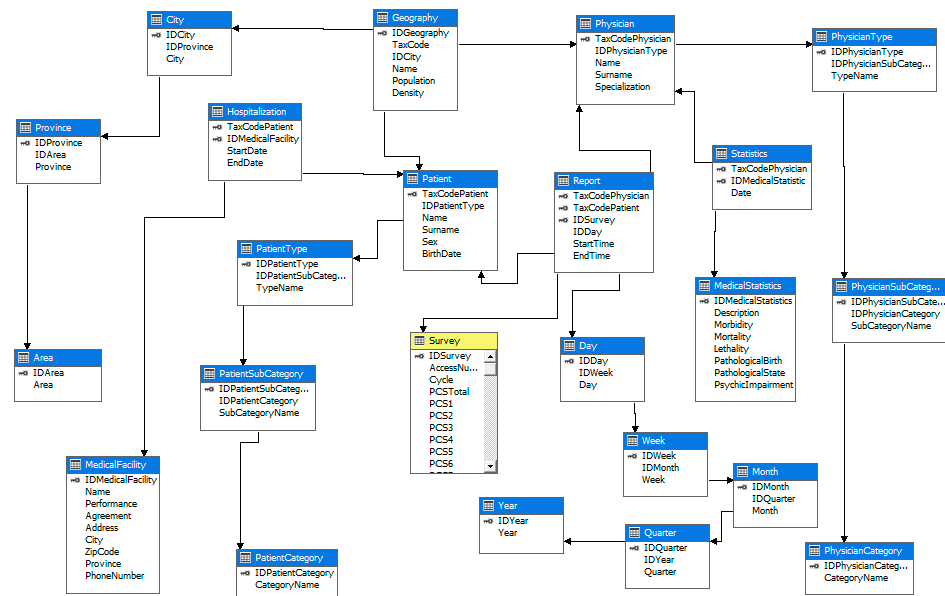


Figure 16. Data Cube *PhysicianCube*.

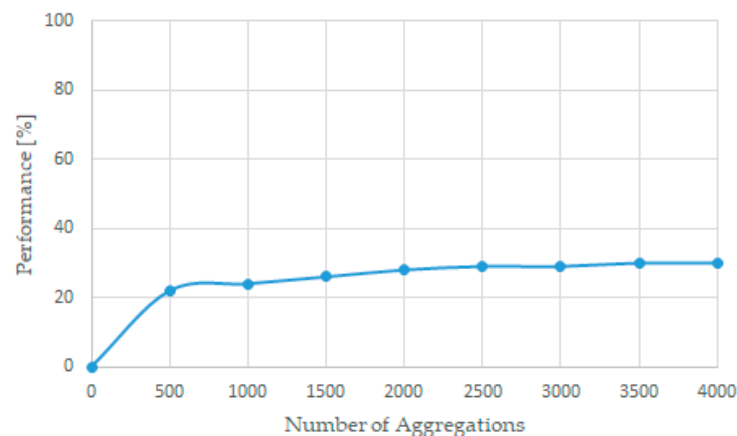


Figure 17. Aggregation Performance Gain on *PhysicianCube*.

Additionally, *MedicalStatistics* plays a pivotal role as an extension of the hierarchy of the dimension *Physician*, thus allowing for analysis of medical statistics and patient records associated with different physicians. This interconnected structure of *PhysicianCube* facilitates a sophisticated analysis of patient medical statistics, enabling users to explore relationships between physicians, patients, and medical outcomes at various levels of granularity.

Crucial information regarding healthcare professionals can be gathered through the data cube *PhysicianCube*. A comprehensive evaluation of *medical well-being* cannot overlook the role of therapists, as their involvement significantly impacts patient outcomes. For statistical and analytical purposes, it is essential to collect detailed data about physicians,

including their *PhysicianType*, *PhysicianCategory*, *PhysicianSubCategory*, and the *Medical-Statistics* they generate, particularly focusing on the *Patient* axis of analysis. Additionally, incorporating geographical details of healthcare professionals, such as their *city of origin*, *province*, and *reference area*, enriches the general *physicians' analysis*. The so-gathered geographical data provides insights into the distribution of physicians and their availability across regions, which is valuable for policy and healthcare planning.

In this context, we delve deeper into the analysis by examining the dimensions of *Physician* and *Geography*. Specifically, the attribute *City* within the hierarchy *Geography* allows us to identify patterns in the geographic distribution of healthcare professionals. This facilitates the identification of cities with the highest and lowest number of physicians. Figure 18 reports the results of the so-conducted aggregation analysis.

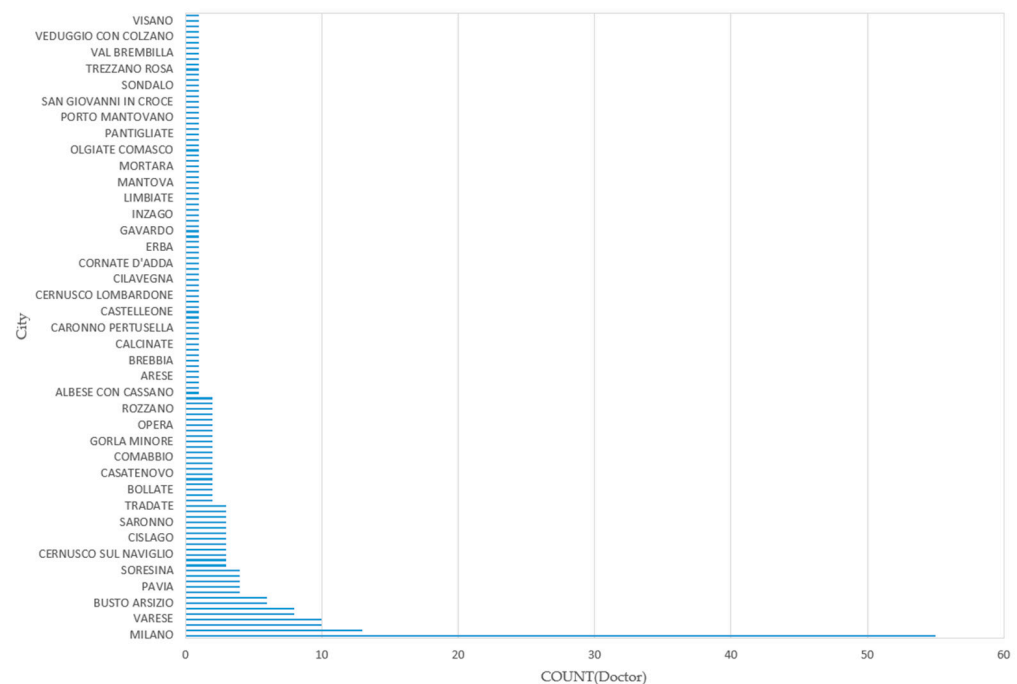


Figure 18. *PatientCube_MDPivoting on PhysicianCube.*

As depicted in Figure 18, *MILANO* emerges as the city with the largest number of physicians, totaling 55 healthcare professionals. Afterward, *VARESE* follows with 13 physicians. Conversely, a group of cities composed of *VISANO*, *VEDUGGIO CON COLZANO*, *VAL BREMBILLA*, *TREZZANO ROSA*, *SONDALO*, *SAN GIOVANNI IN CROCE*, *PORTO MANTOVANO*, *PANTIGLIATE*, *OLGIATE COMASCO*, *MORTARA*, *MANTOVA*, *LIMBIATE*, *INZAGO*, *GAVARDO*, *ERBA*, *CORNATE D'ADDA*, *CILAVEGNA*, *CERNUSCO LOMBARDO*, *CASTELLEONE*, *CARONNO PERTUSELLA*, *CALCINATE*, *BREBBIA*, *ARESE* and *ALBESE CON CASSANO*, with each one reporting an equal and notably low number of physicians, with only 1 physician per city. These findings provide a clear view of physician distribution and highlight areas with potentially underserved healthcare populations.

Multidimensional Clustering analysis is carried out as well. In Figure 19, *Cluster* refers to *Population of PhysicianCube*. *Cluster 1* is the largest in terms of population, equal to 15. Followed, in descending order, by *Cluster 2* with 15, *Cluster 6* with 9, *Cluster 8* with 8, *Cluster 3* with 7, *Cluster 5* with 7, *Cluster 4* with 4, *Cluster 7* with 3, and *Cluster 9* with 3.

Figure 20 shows the cluster profiles for the clustering of Figure 19, which summarizes in a short form the various variables shown in previous clustering diagrams.

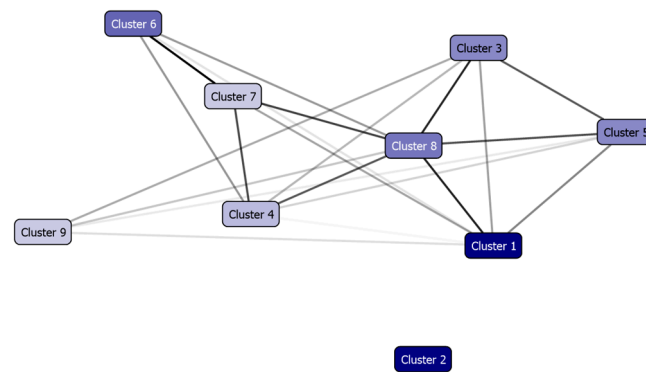


Figure 19. *PhysicianCube_MDClustering on PhysicianCube—Cluster Variable: Population.*

Variables	States	Population... Size: 71	Cluster 2 Size: 15	Cluster 1 Size: 15	Cluster 6 Size: 9	Cluster 8 Size: 8	Cluster 5 Size: 7	Cluster 3 Size: 7	Cluster 4 Size: 4	Cluster 9 Size: 3	Cluster 7 Size: 3
Category Name	- Physiatrist - Neurologist - Rehabilitation - Oncologist - Other										
Specialization	- Medical surgeon - Physiotherapist - missing										
Sub Category Name	- High - Medium - Low - missing										
Type Name	- Doctor - Head of the hos - missing										

Figure 20. *PhysicianCube_MDClustering on PhysicianCube—Cluster Profiles.*

Figure 21 shows clustering discrimination analysis for two different clusters. Here, in Figure 21, *Cluster 1* and *Cluster 9* are compared by highlighting differences. Therefore, with regard to the *count* of physicians based on *SubCategoryName*, the distribution is as follows: about 100% of *SubCategoryName High* in favor of *Cluster 1* compared to 44.989% of the *SubCategoryName Low* rate in *Cluster 9*. Moreover, regarding physicians based on their *CategoryName*, we have a distribution of 31.998% in favor of *Cluster 1* category *Physiatrist* in comparison to *Rehabilitation* having 22.904% and *Neurologist* being only 2.605% in *Cluster 9*. Finally, for the *Specialization* of the physicians, the distribution is equal between *Cluster 1* and *Cluster 9*, being 22.904%.

Variables	Values	Favors Cluster 1	Favors Cluster 9
Sub Category Name	High		
Sub Category Name	Low		
Category Name	Physiatrist		
Specialization	Physiotherapist		
Specialization	Medical surgeon		
Category Name	Rehabilitation		
Sub Category Name	Medium		
Category Name	Neurologist		

Figure 21. *PhysicianCube_MDClustering on PhysicianCube—Cluster Discrimination between Cluster 1 and Cluster 2.*

The clustering analysis on *PhysicianCube* reveals distinct patterns of physician distribution, specialization, and care delivery models (e.g., [67–69]). We focus here on the top three clinically meaningful clusters, which are the following ones:

- *Cluster 1* (*Population* = 15). This cluster is dominated by physicians with *high SubCategoryName* (approximately 100%) and a *specialization distribution* primarily in *Physiatrist* (31.998%) and *Rehabilitation* (22.904%) specialties. Physicians in this cluster are concentrated in urban areas, particularly *MILANO* (which has the highest number of

physicians overall, $n = 55$). From a clinical management perspective, this cluster represents *tertiary care centers* and *specialized rehabilitation facilities* that manage complex chronic pain cases requiring multidisciplinary input. The high concentration of physiatrists and rehabilitation specialists suggests that patients in areas served by this cluster have access to comprehensive rehabilitative care, which is associated with better functional outcomes in chronic pain management.

- *Cluster 2 (Population = 15)*. This cluster exhibits a more balanced distribution of physician specializations, including *Neurologist* (22.904%) alongside physiatrists and rehabilitation specialists. Physicians in this cluster are more evenly distributed across the Lombardy region, including suburban and semi-urban areas. From a clinical perspective, this cluster represents *secondary care centers* and *community-based pain management services* that serve as the first line of intervention for patients with chronic pain. The presence of neurologists alongside physiatrists suggests that this cluster manages a broader range of pain conditions, including those with neurological components (e.g., neuropathic pain), making it a critical component of the regional pain management network.
- *Cluster 3 (Population = 7)*. This cluster is characterized by a higher proportion of physicians with *low SubCategoryName* (44.989%) and *limited specialization diversity*, with rehabilitation specialists being the predominant category. Physicians in this cluster are located in *rural areas* and smaller cities. From a clinical management perspective, this cluster represents *underserved areas* with limited access to specialized chronic pain care. The predominance of general rehabilitation specialists suggests that patients in these areas may have limited access to multimodal pain management approaches, potentially leading to suboptimal pain control and higher reliance on primary care or emergency services. This finding has important *policy implications*, highlighting the need for targeted workforce development and telemedicine initiatives to improve pain management access in rural areas.

The physician cluster profiles reveal significant regional disparities in access to specialized chronic pain care. The concentration of specialists in urban areas (Cluster 1) and the limited specialization in rural areas (Cluster 9) suggest a need for *geographically targeted workforce development* and the expansion of *tele-pain services* to reach underserved populations. Furthermore, the specialization distribution across clusters highlights the importance of *multidisciplinary pain management models* that integrate physiatry, neurology, and rehabilitation expertise. Healthcare policymakers can use these insights to plan resource allocation, training programs, and service delivery models that address regional gaps in chronic pain care.

Following that, we perform the Multidimensional Regression analysis over *PhysicianCube* in order to predict values of the measures of pain on the basis of *PainRELifedb* historical values. Figure 22 shows the various measures, each one with a different color associated with it.

Figure 23 shows predictive analysis based on the previously selected measures. The developed analysis reports percentages of the last current values, represented by the last non-dashed values, and percentages of the last predicted values. In more detail, the percentages are the following:

- *Lethality (blue)*: current value –12%, last predicted value 48%;
- *Morbidity (red)*: current value –50%, last predicted value 27%;
- *Mortality (green)*: current value –42%, last predicted value –16%;
- *Pathological Birth (purple)*: current value –85%, last predicted value –43%;
- *Pathological State (light blue)*: current value –62%, last predicted value –51%;
- *Psychic Impairment (orange)*: current value –12%, last predicted value –7%.



Figure 22. *PhysicianCube_MDRegression on PhysicianCube—Measures.*

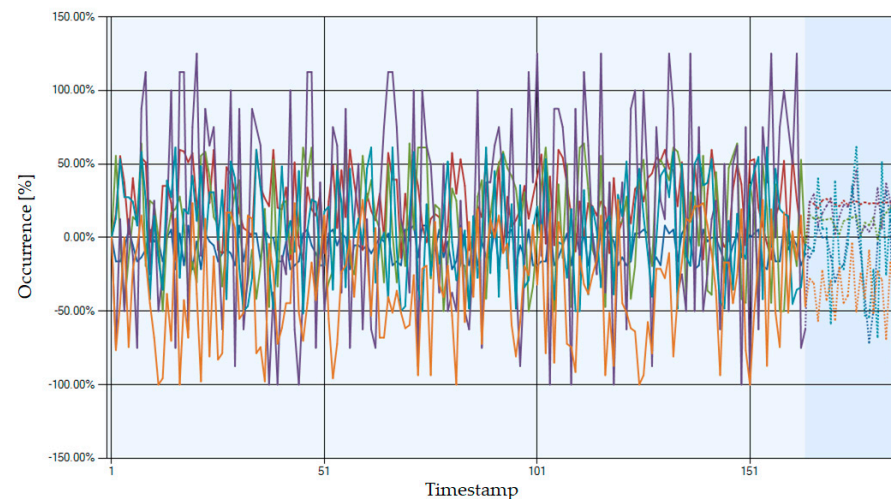


Figure 23. *PhysicianCube_MDRegression on PhysicianCube based on the Measure reported in Figure 22.*

Finally, *PathologyCube*, as depicted in Figure 24, is meticulously designed to efficiently store, organize, and manage data specifically related to patient suffering, *pathology*, *genetics*, and *clinical physiology*. The core architecture of this data cube centers on the fact table *Survey*, which is intricately connected to several key dimensions, including *Pathology*, *ClinicalPhysiology*, *Genetics*, and *Report*. These dimensions are integral to providing a comprehensive and multi-faceted representation of data, thereby enabling detailed analyses from various perspectives.

To further enrich the depth and precision of patient *pathology* suffering data analysis, the *Patient* dimension is subdivided into multiple hierarchical levels. These levels offer a layered approach to classification and segmentation, allowing for a more granular examination of patient data. The hierarchical structures within the *Patient* dimension include *PathologyType*, *PathologySubCategory*, and *PathologyCategory*. Together, these hierarchical levels enable a more nuanced and precise categorization of patients, based on specific clinical and demographic criteria. This stratified approach provides a more detailed view of the ways in which patients experience and are affected by various *pathologies*.

The hierarchical organization within the *Patient* dimension plays a crucial role in ensuring that *PathologyCube* can efficiently process and segment data related to *patient suffering*, *pathologies*, *genetics*, and *clinical physiology* based on the specific roles and specialties of physicians. This structure allows healthcare professionals to conduct sophisticated analyses by focusing on different layers of data, facilitating more personalized treatment strategies and improving the accuracy of clinical decision-making.

Furthermore, the interconnectedness of these dimensions within the *PathologyCube* framework provides an opportunity for multidimensional data analysis, where interactions between *pathology*, *genetics*, and *clinical physiology* can be explored in-depth. This multi-

dimensionality supports complex queries and enables researchers to uncover patterns and correlations across various domains of healthcare data. As a result, the system promotes more efficient clinical workflows and supports ongoing research initiatives aimed at understanding the complex interplay of factors influencing patient outcomes. By enabling a comprehensive approach to data analysis, *PathologyCube* is positioned to significantly advance both clinical practice and patient care.

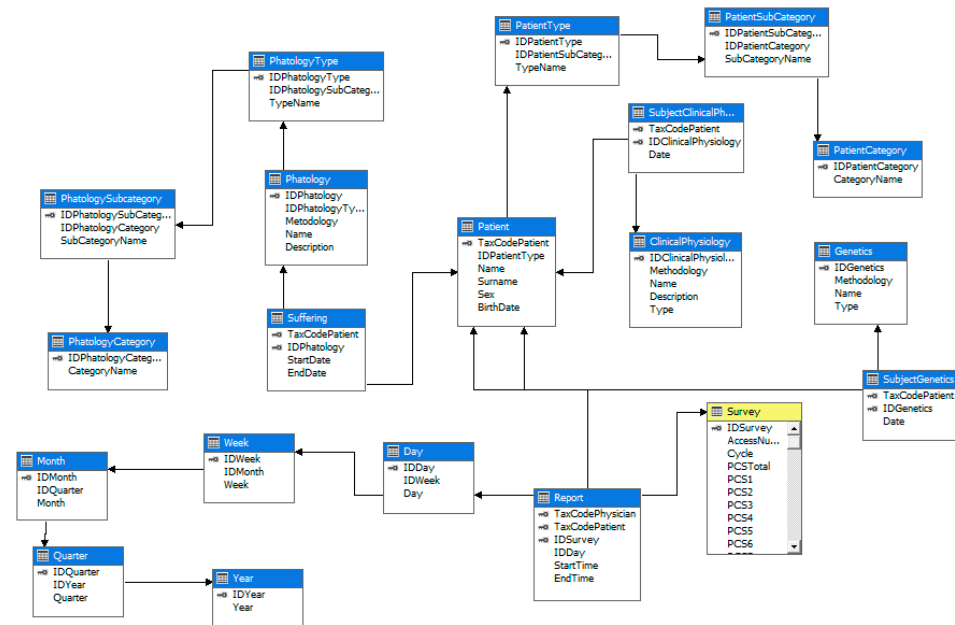


Figure 24. Data Cube *PathologyCube*.

Figure 25 shows the result plot of the query response time optimization based on the number of computed aggregations on *PathologyCube*, where we obtain up to a 24% performance gain in query response time against the pre-computed aggregations.

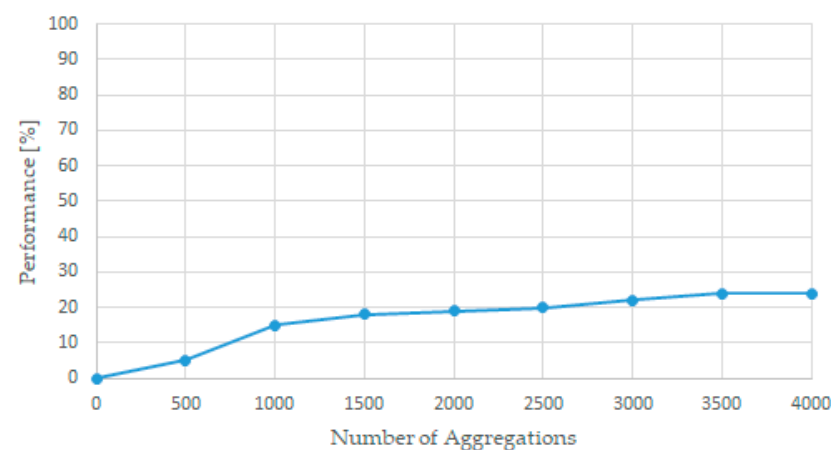


Figure 25. Aggregation Performance Gain on *PathologyCube*.

We can see that the maximum gain is obtained when using the number of aggregations of 3500 and greater. The *Pathology* dimension primarily focuses on the study of *pathology* based on patient data, placing significant emphasis on factors such as *genetics* and *patient physiology*. This dimension serves as a cornerstone in the *PathologyCube* framework, enabling the analysis of various *pathologies* by integrating genetic and physiological information from patient records. In a similar manner to the *Genetics* dimension, which is a critical component of the *Patient* dimension hierarchy, the *Pathology* dimension offers an in-depth

exploration of how specific *pathologies* manifest and progress based on unique genetic and physiological profiles.

The integration of *genetics* within the *Pathology* dimension allows for a more comprehensive and multifactorial approach to understanding patient *pathologies*. Through detailed analysis of *patient records*, users can examine how different *genetic* variations influence the onset and progression of diseases, while also taking into account the *physiological characteristics* that may either exacerbate or mitigate the effects of these conditions. This multifaceted approach provides valuable insights into the intricate relationship between *genetic predispositions*, *physiological traits*, and disease outcomes.

Furthermore, the interconnected structure of *PathologyCube* facilitates sophisticated analyses by enabling users to explore complex relationships between *pathology*, *genetics*, and *clinical physiology* across multiple levels of granularity. By linking these dimensions, the system allows for an enhanced understanding of how specific *genetic* mutations or variations in *physiological characteristics* may influence patient susceptibility to different *pathologies* and how these factors, in turn, impact medical outcomes. This depth of analysis supports personalized medicine, where treatment plans can be tailored to the unique genetic and physiological profiles of individual patients.

At a more granular level, the *Pathology* dimension allows users to examine the effects of specific *pathologies* on various subgroups of patients, taking into account factors such as age, sex, ethnicity, and family history. This level of segmentation enables the identification of potential risk factors and the development of targeted therapeutic strategies. Additionally, the cross-dimensional relationships facilitated by *PathologyCube* enable researchers and healthcare providers to investigate how genetic predispositions interact with environmental factors and clinical conditions to influence the course of disease, providing a holistic view of patient health and pathology.

We aim to gather crucial information regarding healthcare professionals through the data cube *PathologyCube*. A comprehensive evaluation of *medical well-being* cannot overlook the role of genetics and patient physiological characteristics, as their involvement significantly impacts patient outcomes.

For analytical purposes, it is essential to collect detailed data about *pathology genitival* and *physiological* factors, including *PathologyType*, *PathologyCategory*, *PathologySubCategory*, *ClinicalPhysiology*, and the *Genetics* they generate, particularly focusing on the *Patient* axis of analysis by classifying patients via *PatientType*, *PatientSubCategory*, and *PatientCategory*. Additionally, incorporating geographical details of patients, such as their *city of origin*, *province*, and *reference area*, enriches the general *patients' analysis*. This geographical data provides insights into the distribution of patient pathologies, genetics, and physiological characteristics and their availability across regions, which is valuable for healthcare analytics.

In this context, we will delve deeper into the analysis by examining the *Pathology* dimension. Specifically, the *Name* attribute within the *Patient* hierarchy will allow us to identify patterns in the distribution of patient pathology. This will facilitate the identification of cities with the highest and lowest number of disease cases. Figure 26 presents a histogram that visually represents the results of this analysis.

As depicted in Figure 26, *Ictus* and *Cancro al seno* emerge as the pathologies with the largest number of patient cases, totaling 13,000 cases. In contrast, *Aterosclerosi* follows with 13 physicians. Conversely, *Aterosclerosi* reports a significantly low number of cases, with only 800.

Multidimensional Clustering analysis is carried out as well. In Figure 27, *Cluster* refers to the *Population* of *PathologyCube*. *Cluster 1* is the largest in terms of population, equal to 3 medical pathology patient cases. Followed, in descending order, by *Cluster 3* with 2,

Cluster 2 with 2, Cluster 5 with 2, Cluster 4 with 2, Cluster 6 with 2, Cluster 8 with 1, Cluster 7 with 1, Cluster 9 with 1, and Cluster 10 with 1 case.

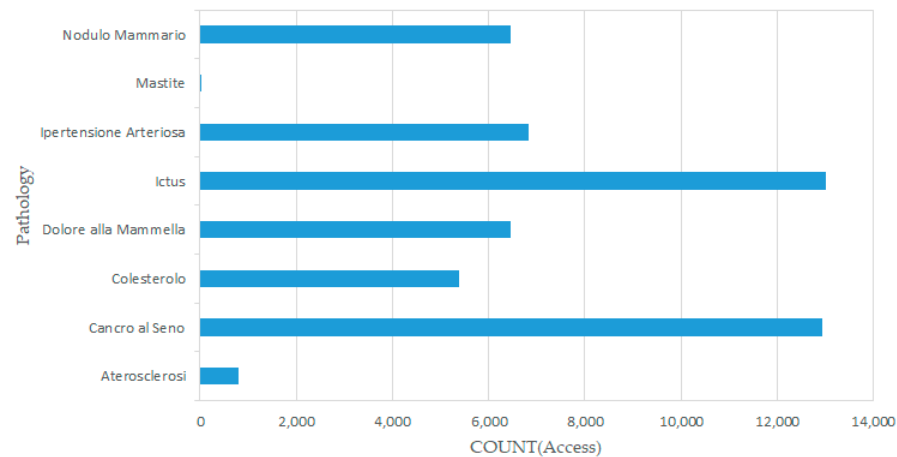


Figure 26. PatientCube_MDPivoting on PathologyCube.

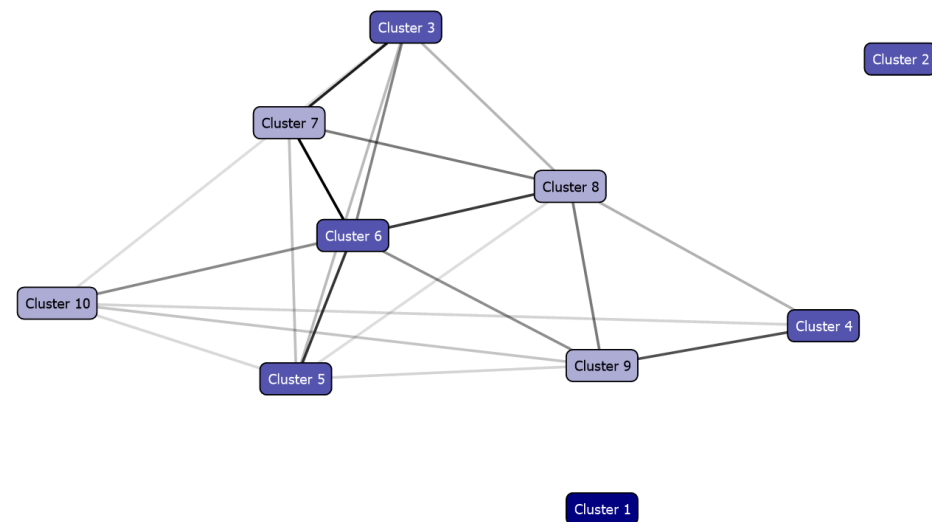


Figure 27. PathologyCube_MDClustering on PathologyCube—Cluster Variable: Population.

Figure 28 shows the *Cluster Profiles* for the clustering of Figure 27, which summarizes in a simplistic and graphically intuitive way data relating to the various variables that the previous cluster diagrams shown have indicated. Figure 29 shows the *Cluster Discrimination* section in which values present in two different *Clusters* are compared. In particular, in the example in Figure 29, Cluster 1 and Cluster 2 are compared, as the differences shown are common to all others.

Relevant differences, therefore, concern the attribute *Pathology Category Name*, whereas regarding Cluster 1, we have about 14.348% of *Genitourinary disorders*, which is less compared to 100% of the *Neoplasms* category in Cluster 2.

The clustering analysis on *PathologyCube* reveals distinct patterns of pathology distribution, genetic factors, and physiological characteristics (e.g., [67–69]). We focus here on the top three clinically meaningful clusters, which are the following ones:

- *Cluster 1* (*Population* = 3). This cluster is characterized by *genitourinary disorders* (14.348%) and a mix of other pathology categories. The relatively small size of this cluster ($n = 3$) suggests that these are *less prevalent pathologies* that require specialized diagnostic and treatment approaches. From a *clinical management perspective*, the presence of *genitourinary disorders* in this cluster highlights the need for *integrated*

care pathways that address both the primary pathology and associated pain symptoms. Patients with genitourinary disorders often experience chronic pelvic pain, which requires a multidisciplinary approach involving urologists, gynecologists, and pain specialists. The clustering of these relatively rare conditions may facilitate the identification of best practices and referral patterns for managing complex, low-prevalence pain conditions.

- **Cluster 2 (Population = 2).** This cluster is exclusively characterized by *neoplasms* (100%), representing patients with *cancer-related chronic pain*. From a clinical management perspective, this cluster aligns with the *breast cancer* subgroup identified in the *PatientCube* analysis (Cluster 2), reinforcing the need for *multimodal pain management approaches* that combine pharmacological interventions with psychological support. The exclusive presence of neoplasms in this cluster suggests that cancer-related pain has distinct pathophysiological and psychological characteristics that differentiate it from other chronic pain conditions. This finding supports the development of *cancer-specific pain management guidelines* that address both the biological drivers of cancer pain (e.g., tumor invasion, nerve compression) and the psychosocial dimensions (e.g., fear of progression, existential distress).
- **Cluster 3 (Population = 2).** This cluster is characterized by a mix of pathology categories, suggesting *complex and multi-morbid pain conditions* that require *integrated, coordinated care approaches*. Patients in this cluster likely have multiple co-existing conditions that complicate pain management. From a clinical perspective, this cluster represents the most challenging patient subgroup, requiring comprehensive, patient-centered care that addresses both pain and underlying comorbidities. This finding highlights the importance of *integrated care models* that bring together primary care, specialty care, and mental health services to address the complex needs of multi-morbid patients with chronic pain.

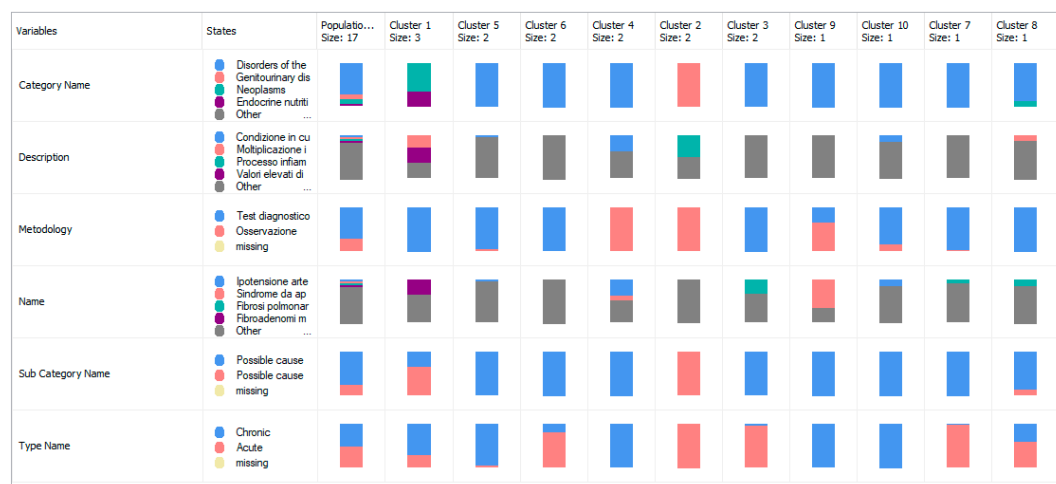


Figure 28. PathologyCube_MDClustering on PathologyCube—Cluster Profiles.

The pathology cluster profiles reveal distinct patterns of disease distribution, with cancer-related pain emerging as a distinct entity that requires specialized, multimodal approaches. The presence of rare pathologies (Cluster 1) and complex multi-morbid conditions (Cluster 3) highlights the need for *personalized care pathways* that consider both the primary pathology and associated comorbidities. Additionally, the clustering of relatively rare conditions may facilitate the development of *clinical decision support systems* that help clinicians identify best practices and referral pathways for specific pathology subgroups.

Variables	Values	Favors Cluster 1	Favors Cluster 2
Metodology	Osservazione		
Category Name	Genitourinary disorders		
Metodology	Test diagnostico		
Type Name	Acute		
Type Name	Chronic		
Category Name	Neoplasms		
Name	Mastite		
Name	Dolore alla mammella		
Description	Processo infiammatorio che interess...		
Description	Dolore diffuso o localizzato alla mam...		

Figure 29. PathologyCube_MDClustering on PathologyCube—Cluster Discrimination between Cluster 1 and Cluster 2.

Following that, we perform the Multidimensional Regression analysis over *PathologyCube* in order to predict values of the measures of pain on the basis of *PainRELifeDB* historical values. Figure 30 shows measures to which a different color is associated.

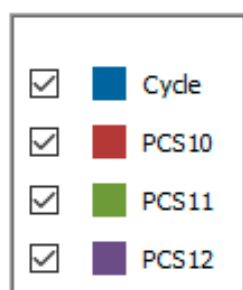


Figure 30. PathologyCube_MDRegression on PathologyCube—Measures.

Figure 31 shows predictive analysis based on the previously selected measures. Analysis reports percentages of the last current values, represented by the last non-dashed values, and percentages of the last predicted values. The percentages are the following:

- *Cycle* (blue): current value −70%, last predicted value 50%;
- *PCS10* (red): current value −30%, last predicted value 0%;
- *PCS11* (green): current value −0%, last predicted value −0%;
- *PCS12 Birth* (purple): current value −20%, last predicted value −28%.

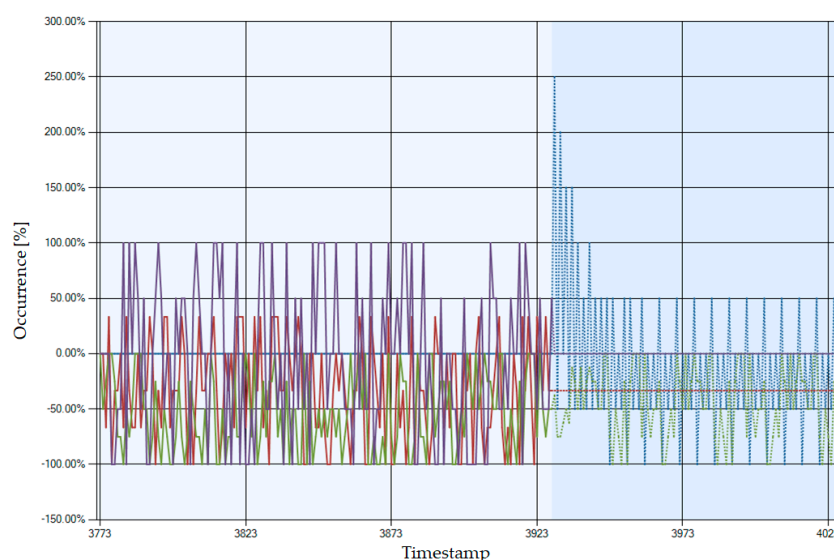


Figure 31. PathologyCube_MDRegression on PathologyCube based on the Measure reported in Figure 30.

6. Discussion and Remarks

In this Section, we provide a discussion of our results as well as some future research directions to improve our proposed framework.

It is important to note that the analyses presented in Figures 10, 18 and 26 are *descriptive and exploratory* in nature (e.g., [71]). They are intended to provide a visual summary of the data distribution and to illustrate the types of aggregation analyses that our framework supports (e.g., identifying cities with the highest/lowest patient counts, physician distributions, or pathology case counts). Formal statistical significance tests (e.g., *chi-square tests* [72]) were not performed on these descriptive visualizations, as they are not designed to test specific hypotheses about population differences but rather to facilitate initial data exploration and pattern discovery. However, future work will extend our framework to incorporate formal hypothesis testing capabilities for comparative subgroup analyses in healthcare analytics.

Moreover, the clustering results, presented in Section 5.2, are *descriptive and exploratory* in nature (e.g., [73]). These results provide a visual summary of the natural groupings within the multidimensional data and illustrate the types of patient, physician, and pathology segmentations that our framework can identify. Furthermore, the clinical profiles derived from these clusters provide face validity by demonstrating that the clusters correspond to clinically meaningful patient subgroups with distinct pain characteristics, psychological profiles, and care utilization patterns. Moving forward, we plan to extend our framework to incorporate formal hypothesis testing capabilities for comparative subgroup analyses in healthcare analytics. This would enable, for example, statistical comparisons of pain outcomes across clusters (e.g., [74]) or tests for significant differences in treatment response rates between identified patient subgroups.

Finally, the percentages reported in the predictive analysis (see Figures 15, 23 and 31) represent the *relative change* in each measure compared to its *historical baseline*. The historical baseline is defined as the *average value of each measure over the entire historical period* available in the *PainRELifeDB* database. Specifically, we have the following:

- The *current value* percentage denotes the relative change in the most recent observed value compared to the historical baseline.
- The *last predicted value* percentage denotes the relative change in the forecasted value at the final time step of the prediction horizon compared to the same historical baseline.

For example, a current value of -34% for BPI A (blue line in Figure 15) indicates that the most recent observed BPI A score is 34% lower than the historical average for that measure. A last predicted value of 67% indicates that the model forecasts a 67% increase relative to the historical baseline at the end of the prediction horizon.

While the experimental results presented in Section 5 demonstrate the effectiveness of our *Cloud-Based Multidimensional Big Data Analytics framework* in a controlled research setting, translating this framework into a production-ready clinical system presents several significant challenges that must be acknowledged and addressed, which are the following:

- **Data Heterogeneity and Integration Complexity:** Our framework operates on a structured, harmonized input derived from the *PainRELifeDB* database. However, real-life healthcare environments are characterized by data fragmentation across multiple, often incompatible systems, including *Electronic Health Records* (EHRs), wearable devices, imaging systems, and laboratory information systems. Integrating these heterogeneous sources while maintaining semantic consistency and temporal alignment poses substantial technical challenges. Future work must incorporate robust *Extract-Transform-Load* (ETL) pipelines (e.g., [75]) and adopt semantic interoperability

standards such as HL7 FHIR to facilitate seamless data ingestion from diverse clinical sources [76].

- **Regulatory and Ethical Compliance:** The processing of PHI in cloud-based environments is subject to strict regulatory frameworks, including GDPR in Europe and HIPAA in the United States (e.g., [24,25]). While our framework incorporates de-identification protocols and data isolation mechanisms (as discussed in Section 2), real-life deployment requires continuous compliance monitoring, regular security audits, and explicit patient consent mechanisms for secondary use of clinical data (e.g., [77]). The evolving nature of data protection legislation demands that any production-ready framework maintain flexibility to adapt to new regulatory requirements. Furthermore, ethical considerations regarding algorithmic bias and fairness must be systematically addressed (e.g., [78]). Future work will explore the integration of privacy-preserving computation techniques, such as federated learning and differential privacy, to further enhance patient data protection while preserving analytical utility.
- **Algorithmic Interpretability and Explainability:** Healthcare practitioners are traditionally trained to make decisions based on interpretable, evidence-based reasoning. Although our *Multidimensional Regression* tool with *Lasso regularization* provides automatic feature selection (thereby enhancing interpretability compared to fully opaque models), and our Multidimensional Clustering visualizations offer intuitive cluster profiling, the underlying ML algorithms (i.e., *K-Means++* and *Multivariate Linear Regression*) may still be perceived as “black boxes” (e.g., [79]) by clinicians who are not familiar with computational analytics. This perception creates a barrier to adoption, as clinical decision-makers must trust the system’s outputs before integrating them into patient care workflows. To address this, future work will incorporate *Explainable AI* (XAI) techniques [80] specifically tailored to multidimensional healthcare data, including the generation of local explanations (e.g., *SHAP* values, *LIME*) for individual predictions and global explanations (e.g., feature importance rankings, decision rules) to help clinicians understand the drivers of model outputs. Recent advances in XAI for smart healthcare environments (e.g., [81]) provide promising directions for bridging this interpretability gap.
- **Scalability to Larger and More Diverse Datasets:** The *PainRELifedB* database, at approximately 2.72 GB with over 1.2 million records, is representative of a regional healthcare repository. However, national or multi-national healthcare systems routinely generate data at the terabyte or petabyte scale, with significantly higher velocity and variety. While our aggregation optimization techniques demonstrated up to 30% performance gains on the current dataset (see Section 5.2), scaling to larger workloads will require further architectural enhancements, including the adoption of distributed computing frameworks (e.g., Apache Spark, Hadoop) and the implementation of incremental learning algorithms that can update models without retraining from scratch (e.g., [82,83]). The exponential growth of healthcare data, driven by the increasing adoption of wearable devices and remote patient monitoring, demands that future versions of our framework incorporate these distributed and streaming capabilities (e.g., [84,85]).
- **Real-Time Decision Support and Latency Guarantees:** While our performance gains (up to 30% query response time improvement) are significant for batch analytics and exploratory OLAP operations, real-time clinical decision support systems often require sub-second latency guarantees for critical queries (e.g., during emergency consultations or acute pain episodes). Achieving such stringent performance targets will require additional optimization strategies, including in-memory caching, streaming analytics pipelines, and predictive pre-fetching of frequently accessed data

cubes. Moreover, the system must provide reliable *Quality-of-Service* (QoS) guarantees (e.g., [86,87]), ensuring that query response times remain consistently low even under peak workload conditions. Future work will investigate the integration of real-time data streaming platforms (e.g., Apache Kafka, Apache Flink) and the development of adaptive caching strategies that dynamically prioritize high-utility queries based on clinical urgency [88].

- **Generalizability:** Our experiments are based on the *Pain-RELife* real-life project, which focuses specifically on patients with chronic pain in the Lombardy region of North Italy. While this provides a rich and realistic case study, the generalizability of our framework to other geographic regions, healthcare systems, and patient populations (e.g., pediatric pain, non-chronic pain conditions) remains to be demonstrated. Future work will involve the application of our framework to diverse healthcare datasets from multiple regions and clinical domains, systematically evaluating its performance across varying data distributions, coding standards, and treatment protocols.

By pursuing these research directions, we aim to transform our current prototype into a robust, production-ready platform that can deliver actionable insights to clinicians and researchers, contributing to the ongoing advancement of patient-centered healthcare analytics and evidence-based chronic pain management.

7. Conclusions and Future Work

Focusing on the emerging healthcare analytics context, this paper has provided practical and experimental results coming from a real-life research project that focuses on *Big Data Analytics over Big Healthcare Data* about chronic pain of patients in real-life hospitals and medical centers in North Italy (Region Lombardy), *Pain-RELife*. In this context, we have provided the innovative Multidimensional Big Data Analytics tools and a reference cloud-based architecture that implements our analytics paradigm, as well as several experimental results that highlight the benefits in both big data analytics and big data predictive analytics.

Future work is mainly focused on dealing with emerging open problems in big data analytics for healthcare (e.g., [89,90]).

Author Contributions: Conceptualization, A.C.; methodology, A.C.; software, A.H.; validation, A.C., and A.H.; formal analysis, A.C.; investigation, A.C.; resources, A.C.; data curation, A.H.; writing—original draft preparation, A.C., and A.H.; writing—review and editing, A.C.; visualization, A.H.; supervision, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was assessed and certified by the ethical board of the European Institute of Oncology (approval number R1597/21-IEO 1701).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Datasets are protected by project constraints.

Acknowledgments: This research is supported by the ICSC National Research Center for High Performance Computing, Big Data and Quantum Computing within the NextGenerationEU program (Project Code: PNRR CN00000013). The authors are grateful to Sara Marceglia, from University of Trieste, Trieste, Italy, for her contributions to the conceptual modeling of the real-life healthcare dataset.

Conflicts of Interest: Authors declare that they have no conflicts of interest. They also declare that they do not have any competing financial interests.

References

1. Cuzzocrea, A. Multidimensional Big Data Analytics over Big Web Knowledge Bases: Models, Issues, Research Trends, and a Reference Architecture. In Proceedings of the 8th IEEE International Conference on Multimedia Big Data, Naples, Italy, 5–7 December 2022; pp. 1–6.
2. Cuzzocrea, A.; Leung, C.K.; MacKinnon, R.K. Mining Constrained Frequent Itemsets from Distributed Uncertain Data. *Future Gener. Comput. Syst.* **2014**, *37*, 117–126. [\[CrossRef\]](#)
3. Cuzzocrea, A. Multidimensional Clustering over Big Data: Models, Issues, Analysis, Emerging Trends. In Proceedings of the 32nd ACM International Conference on Scientific and Statistical Database Management, Vienna, Austria, 7–9 July 2020; art. 32.
4. Cuzzocrea, A. Compressing Big OLAP Data Cubes in Big Data Analytics Systems: New Paradigms and Future Research Perspectives. In Proceedings of the 19th International Conference on Smart Business Technologies, Lisbon, Portugal, 14–16 July 2022; pp. 156–175.
5. Olken, F.; Rotem, D. Random Sampling from Databases: A Survey. *Stat. Comput.* **1995**, *5*, 25–42. [\[CrossRef\]](#)
6. Russom, P. Big Data Analytics. *TDWI Best Pract. Rep. Fourth Quart.* **2011**, *19*, 1–34.
7. Marshall, E.A.; Qanungo, S.; Obeid, J.S.; Oates, J.C.; Habrat, M.L.; Warren, R.W.; Lenert, L.A. Qualitative Analysis of Responses to a Questionnaire via an EHR Patient Portal. In Proceedings of the 2015 American Medical Informatics Association Annual Symposium, San Francisco, CA, USA, 14–18 November 2015.
8. Poon, L.K.M.; Zhang, N.L.; Liu, T.; Liu, A.H. Model-Based Clustering of High-Dimensional Data: Variable Selection versus Facet Determination. *Int. J. Approx. Reason.* **2013**, *54*, 196–215. [\[CrossRef\]](#)
9. Chen, Y.; Dong, G.; Han, J.; Wah, B.W.; Wang, J. Multi-Dimensional Regression Analysis of Time-Series Data Streams. In Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, 20–23 August 2002; pp. 323–334.
10. Ristevski, B.; Chen, M. Big Data Analytics in Medicine and Healthcare. *J. Integr. Bioinform.* **2018**, *15*, 20170030. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Belle, A.; Thiagarajan, R.; Soroushmehr, S.M.; Navidi, F.; Beard, D.A.; Najarian, K. Big Data Analytics in Healthcare. *BioMed Res. Int.* **2015**, *2015*, 370194. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Cuzzocrea, A.; Hafsaoui, A.; Gallo, C. From Theory to Practice of Multidimensional Big Data Analytics over Big Healthcare Data: A Real-Life Case Study. In Proceedings of the 28th International Database Engineered Applications Symposium, Bayonne, France, 26–29 August 2024; pp. 156–171.
13. Cuzzocrea, A. Aggregation and Multidimensional Analysis of Big Data for Large-Scale Scientific Applications: Models, Issues, Analytics, and Beyond. In Proceedings of the 27th ACM International Conference on Scientific and Statistical Database Management, La Jolla, CA, USA, 29 June–1 July 2015; art. 23.
14. Cuzzocrea, A. Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures. In Proceedings of the 37th Annual IEEE Computer Software and Applications Conference, Kyoto, Japan, 22–26 July 2013; pp. 481–483.
15. Beuscart-Zéphir, M.C.; Anceaux, F.; Menu, H.; Guerlinger, S.; Watbled, L.; Evrard, F. User-Centred, Multidimensional Assessment Method of Clinical Information Systems: A Case-Study in Anaesthesiology. *Int. J. Med. Inform.* **2005**, *74*, 179–189. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Park, E.H.; Kim, J.; Wiles, L.L. The Role of Collectivism and Moderating Effect of IT Proficiency on Intention to Disclose Protected Health Information. *Inf. Technol. Manag.* **2023**, *24*, 177–193. [\[PubMed\]](#)
17. Singh, T.; Di Gangi, P.M.; Johnston, A.C.; Bott, G.J.; Lowry, P.B. Understanding the Influence of Data Breaches on Patients' Willingness to Share Protected Health Information: A Mixed Methods Study of a Construals Privacy Calculus Perspective. *Inf. Syst. Front.* **2026**, *28*, 751–776. [\[CrossRef\]](#)
18. Cuzzocrea, A. Scalable Distributed Pivot Analysis over Massive Big Data: Models, Paradigms, New Advancements. In Proceedings of the 20th IEEE International Conference on Data Mining Workshops, Sorrento, Italy, 17–20 November 2020; pp. 696–700.
19. Li, J.; Wang, S.; Rudinac, S.; Osseyran, A. High-Performance Computing in Healthcare: An Automatic Literature Analysis Perspective. *J. Big Data* **2024**, *11*, 61. [\[CrossRef\]](#)
20. Cuzzocrea, A. Innovative Paradigms for Supporting Privacy-Preserving Multidimensional Big Healthcare Data Management and Analytics: The Case of the EU H2020 QUALITOP Research Project. In Proceedings of the 20th International Semantic Web Conference, Online, 24–28 October 2021; pp. 1–7.
21. Wan, S.J.; Wong, S.K.M.; Prusinkiewicz, P. An Algorithm for Multidimensional Data Clustering. *ACM Trans. Math. Softw.* **1988**, *14*, 153–162. [\[CrossRef\]](#)
22. Thakur, A.; Ranga, V.; Agarwal, R. ABHealChain: Enhancing Privacy and Security in Healthcare Data Sharing Through Hyperledger Fabric and Attribute-Based Access Control. *Concurr. Comput. Pract. Exp.* **2026**, *38*, e70592. [\[CrossRef\]](#)
23. Kumar, S.; Shaw, D.K. Efficient Access Requests Management for Healthcare Data with Security and Privacy-Preserving. *Expert Syst. Appl.* **2025**, *267*, 126194. [\[CrossRef\]](#)

24. Mantelero, A. Personal Data for Decisional Purposes in the Age of Analytics: From an Individual to a Collective Dimension of Data Protection. *Comput. Law Secur. Rev.* **2016**, *32*, 238–255. [\[CrossRef\]](#)
25. Kolasa, K.; Ken Redekop, W.; Berler, A.; Zah, V.; Asche, C.V. Future of Data Analytics in the Era of the General Data Protection Regulation in Europe. *PharmacoEconomics* **2020**, *38*, 1021–1029. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Pawar, P.P.; Kumar, D.; Meesala, M.K.; Pareek, P.K.; Addula, S.R.; Shwetha, K.S. A Patient-Centric Blockchain Framework for Transparent and Secure Medical Data Sharing using Modified AES. In Proceedings of the 2024 International Conference on Integrated Intelligence and Communication Systems, Kalaburagi, India, 22–23 November 2024; pp. 1–7.
27. Lastre, J.K.; Ko, Y.; Kwon, H.; You, I. Evaluating Transport Layer Security 1.3 Optimization Strategies for 5G Cross-Border Roaming: A Comprehensive Security and Performance Analysis. *Sensors* **2025**, *25*, 6144. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Sandhu, R.S. Role-Based Access Control. *Adv. Comput.* **1998**, *46*, 237–286. [\[CrossRef\]](#)
29. Androulaki, E.; Barger, A.; Bortnikov, V.; Cachin, C.; Christidis, K.; De Caro, A.; Enyeart, D.; Ferris, C.; Laventman, G.; Manevich, Y.; et al. Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. In Proceedings of the 13th ACM EuroSys Conference, Porto, Portugal, 23–26 April 2018; pp. 1–15.
30. Xiong, H.; Zhang, H.; Sun, J. Attribute-Based Privacy-Preserving Data Sharing for Dynamic Groups in Cloud Computing. *IEEE Syst. J.* **2018**, *13*, 2739–2750.
31. Yin, L.; Feng, J.; Xun, H.; Sun, Z.; Cheng, X. A Privacy-Preserving Federated Learning for Multiparty Data Sharing in Social IoTs. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 2706–2718. [\[CrossRef\]](#)
32. Liu, Y.; Liu, P.; Jing, W.; Song, H.H. PD2S: A Privacy-Preserving Differentiated Data Sharing Scheme Based on Blockchain and Federated Learning. *IEEE Internet Things J.* **2023**, *10*, 21489–21501. [\[CrossRef\]](#)
33. Sullivan, M.J.; Bishop, S.R.; Pivik, J. The Pain Catastrophizing Scale: Development and Validation. *Psychol. Assess.* **1995**, *7*, 524–532. [\[CrossRef\]](#)
34. Cleeland, C.S.; Ryan, K.M. Pain Assessment: Global Use of the Brief Pain Inventory. *Ann. Acad. Med.* **1994**, *23*, 129–138.
35. Farrar, J.T.; Young, J.P., Jr.; LaMoreaux, L.; Werth, J.L.; Poole, R.M. Clinical Importance of Changes in Chronic Pain Intensity Measured on an 11-Point Numerical Pain Rating Scale. *Pain* **2001**, *94*, 149–158. [\[CrossRef\]](#) [\[PubMed\]](#)
36. McHorney, C.A.; Ware, J.E., Jr.; Raczek, A.E. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs. *Med. Care* **1993**, *31*, 247–263. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Cuzzocrea, A.; Soufargi, S. Privacy-Preserving Big Hierarchical Data Analytics via Co-Occurrence Analysis. In Proceedings of the 13th International Conference on Data Science, Technology and Applications, Dijon, France, 9–11 July 2024; pp. 93–103.
38. Ordonez, C.; Chen, Z.; Cuzzocrea, A.; García-García, J. An Intelligent Visual Big Data Analytics Framework for Supporting Interactive Exploration and Visualization of Big OLAP Cubes. In Proceedings of the 24th IEEE International Conference on Information Visualisation, Melbourne, Australia, 7–11 September 2020; pp. 421–427.
39. Cuzzocrea, A.; Moussa, R. A Cloud-Based Framework for Supporting Effective and Efficient OLAP in Big Data Environments. In Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, Chicago, IL, USA, 26–29 May 2014; pp. 680–684.
40. Akter, M.S.; Shahriar, H.; Cuzzocrea, A.; Wu, F.; Rodriguez-Cardenas, J. A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data. In Proceedings of the 2023 IEEE International Conference on Big Data, Sorrento, Italy, 15–18 December 2023; pp. 5418–5427.
41. Cuzzocrea, A.; De Bock, G.H.; Maas, W.J.; Soufargi, S.; Hafsaoui, A. An Innovative Big Temporal Data Analytics Technique over Real-Life Healthcare Datasets: The F-TBDA Approach. In Proceedings of the 32nd Symposium of Advanced Database Systems, Villasimius, Italy, 23–26 June 2024; pp. 112–121.
42. Cuzzocrea, A.; Saccà, D. Balancing Accuracy and Privacy of OLAP Aggregations on Data Cubes. In Proceedings of the 13th ACM International Workshop on Data Warehousing and OLAP, Toronto, ON, Canada, 30 October 2010; pp. 93–98.
43. Cuzzocrea, A.; Russo, V.; Saccà, D. A Robust Sampling-Based Framework for Privacy Preserving OLAP. In Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy, 2–5 September 2008; pp. 97–114.
44. Morris, K.J.; Egan, S.D.; Linsangan, J.L.; Leung, C.K.; Cuzzocrea, A.; Hoi, C.S.H. Token-Based Adaptive Time-Series Prediction by Ensembling Linear and Non-linear Estimators: A Machine Learning Approach for Predictive Analytics on Big Stock Data. In Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA, 17–20 December 2018; pp. 1486–1491.
45. Leung, C.K.; Elias, J.D.; Minuk, S.M.; De Jesus, A.R.R.; Cuzzocrea, A. An Innovative Fuzzy Logic-Based Machine Learning Algorithm for Supporting Predictive Analytics on Big Transportation Data. In Proceedings of the 29th IEEE International Conference on Fuzzy Systems, Glasgow, UK, 19–24 July 2020; pp. 1–8.
46. Xu, W.; Souly, N.; Brahma, P.P. Reliability of GAN Generated Data to Train and Validate Perception Systems for Autonomous Vehicles. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021; pp. 171–180.
47. Krippendorff, K. *The Reliability of Generating Data*; Chapman & Hall: London, UK; CRC Press: Boca Raton, FL, USA, 2022.

48. Shahee, S.A.; Ananthakumar, U. Synthetic Sampling Approach Based on Model-Based Clustering for Imbalanced Data. *Int. J. Artif. Intell. Soft Comput.* **2018**, *6*, 348–364. [[CrossRef](#)]
49. Liu, C.L.; Hsieh, P.Y. Model-Based Synthetic Sampling for Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1543–1556.
50. Schröer, G.; Trenkler, D. Exact and Randomization Distributions of Kolmogorov-Smirnov Tests Two or Three Samples. *Comput. Stat. Data Anal.* **1995**, *20*, 185–202. [[CrossRef](#)]
51. Al-Labadi, L.; Zarepour, M. Two-Sample Kolmogorov-Smirnov Test using a Bayesian Nonparametric Approach. *Math. Methods Stat.* **2017**, *26*, 212–225. [[CrossRef](#)]
52. Marcegaglia, S.R.; Manzelli, V.; D'Antrassi, P.; Cuzzocrea, A.; Caruso, A.; Prenassi, M.; Savino, C.; De Pippo, M.T.; Lenzi, S.; Ferrucci, R.; et al. PainRE-Life: A FHIR Based HUB for the Management and Support of Patients with Chronic Pain. In Proceedings of the 2021 American Medical Informatics Association Annual Symposium, San Diego, CA, USA, 30 October 2021.
53. Lee, C.; Luo, Z.; Ngiam, K.Y.; Zhang, M.; Zheng, K.; Chen, G.; Ooi, B.C.; Yip, W.L.J. Big Healthcare Data Analytics: Challenges and Applications. In *Handbook of Large-Scale Distributed Computing in Smart Healthcare*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 11–41.
54. Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035.
55. Trebuna, P.; Halcinová, J.; Fil'o, M.; Markovic, J. The Importance of Normalization and Standardization in the Process of Clustering. In Proceedings of the 12th IEEE International Symposium on Applied Machine Intelligence and Informatics, Herl'any, Slovakia, 23–25 January 2014; pp. 381–385.
56. Schubert, E. Stop Using the Elbow Criterion for K-Means and how to Choose the Number of Clusters Instead. *SIGKDD Explor.* **2023**, *25*, 36–42. [[CrossRef](#)]
57. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
58. Del Águila, M.R.; Benítez-Parejo, N. Simple Linear and Multivariate Regression Models. *Allergol. Immunopathol.* **2011**, *39*, 159–173. [[CrossRef](#)] [[PubMed](#)]
59. Vidaurre, D.; Bielza, C.; Larranaga, P. A Survey of L1 Regression. *Int. Stat. Rev.* **2013**, *81*, 361–387. [[CrossRef](#)]
60. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared is more Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [[CrossRef](#)] [[PubMed](#)]
61. Gao, J. R-Squared (R²)—How Much Variation is Explained? *Res. Methods Med. Health Sci.* **2024**, *5*, 104–109. [[CrossRef](#)]
62. Olawoyin, A.M.; Leung, C.K.; Cuzzocrea, A. Privacy-Preserving Publishing and Visualization of Spatial-Temporal Information. In Proceedings of the 2021 IEEE International Conference on Big Data, Orlando, FL, USA, 15–18 December 2021; pp. 5420–5429.
63. Gray, J.; Chaudhuri, S.; Bosworth, A.; Layman, A.; Reichart, D.; Venkatrao, M.; Pellow, F.; Pirahesh, H. Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals. *Data Min. Knowl. Discov.* **1997**, *1*, 29–53. [[CrossRef](#)]
64. Cuzzocrea, A.; Folino, F.; Guarascio, M.; Pontieri, L. A Predictive Learning Framework for Monitoring Aggregated Performance Indicators over Business Process Events. In Proceedings of the 22nd ACM International Database Engineering and Applications Symposium, Villa San Giovanni, Italy, 18–20 June 2018; pp. 165–174.
65. Cuzzocrea, A. An OLAM-Based Framework for Complex Knowledge Pattern Discovery in Distributed-and-Heterogeneous-Data-Sources and Cooperative Information Systems. In Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery, Regensburg, Germany, 3–7 September 2007; pp. 181–198.
66. Cuzzocrea, A. Computing and Mining ClustCube Cubes Efficiently. In Proceedings of the 19th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, 19–22 May 2015; pp. 146–161.
67. Turk, D.C.; Okifuji, A. Psychological Factors in Chronic Pain: Evolution and Revolution. *J. Consult. Clin. Psychol.* **2002**, *70*, 678–690. [[CrossRef](#)] [[PubMed](#)]
68. Keefe, F.J.; Rumble, M.E.; Scipio, C.D.; Giordano, L.A.; Perri, L.M. Psychological Aspects of Persistent Pain: Current State of the Science. *J. Pain* **2004**, *5*, 195–211. [[CrossRef](#)] [[PubMed](#)]
69. Gatchel, R.J.; Peng, Y.B.; Peters, M.L.; Fuchs, P.N.; Turk, D.C. The Biopsychosocial Approach to Chronic Pain: Scientific Advances and Future Directions. *Psychol. Bull.* **2007**, *133*, 581–624. [[CrossRef](#)] [[PubMed](#)]
70. Localio, A.R.; Goodman, S. Beyond the Usual Prediction Accuracy Metrics: Reporting Results for Clinical Decision Making. *Ann. Intern. Med.* **2012**, *157*, 294–295. [[CrossRef](#)] [[PubMed](#)]
71. Morfonios, K.; Konakas, S.; Ioannidis, Y.; Kotsis, N. ROLAP Implementations of the Data Cube. *ACM Comput. Surv.* **2007**, *39*, 12. [[CrossRef](#)]
72. Lewis, D.; Burke, C.J. The Use and Misuse of the Chi-Square Test. *Psychol. Bull.* **1949**, *46*, 433–489. [[CrossRef](#)] [[PubMed](#)]
73. Hosseini Rad, M.; Abdolrazzagah-Nezhad, M. A New Hybridization of DBSCAN and Fuzzy Earthworm Optimization Algorithm for Data Cube Clustering. *Soft Comput.* **2020**, *24*, 15529–15549. [[CrossRef](#)]
74. Rodriguez, M.Z.; Comin, C.H.; Casanova, D.; Bruno, O.M.; Amancio, D.R.; Costa, L.D.F.; Rodrigues, F.A. Clustering Algorithms: A Comparative Approach. *PLoS ONE* **2019**, *14*, e0210236. [[CrossRef](#)]

75. Zarate, G.; Lopez Osa, M.J.; Torre-Bastida, A.I.; Iturraspe, U.; Arjona, J.; Navarro, B.; Gimeno, A. Evolution of Extract-Transform-Load (ETL) Processes Towards Data Product Pipelines. In Proceedings of the 4th ACM Eclipse Security, AI, Architecture and Modelling Conference on Data Space, Mainz, Germany, 22 October 2024; pp. 25–32.
76. Solar, M.; Castaneda, V.; Nanculef, R.; Dombrowskaia, L.; Araya, M. A Data Ingestion Procedure Towards a Medical Images Repository. *Sensors* **2024**, *24*, 4985. [[CrossRef](#)] [[PubMed](#)]
77. Goldschmitt, M.; Gleim, P.; Mandelartz, S.; Kellmeyer, P.; Rigotti, T. Digitalizing Informed Consent in Healthcare: A Scoping Review. *BMC Health Serv. Res.* **2025**, *25*, 893. [[CrossRef](#)] [[PubMed](#)]
78. Kordzadeh, N.; Ghasemaghahi, M. Algorithmic Bias: Review, Synthesis, and Future Research Directions. *Eur. J. Inf. Syst.* **2022**, *31*, 388–409. [[CrossRef](#)]
79. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [[CrossRef](#)]
80. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93. [[CrossRef](#)]
81. Vani, M.S.; Sudhakar, R.V.; Mahendar, A.; Ledalla, S.; Radha, M.; Sunitha, M. Personalized Health Monitoring using Explainable AI: Bridging Trust in Predictive Healthcare. *Sci. Rep.* **2025**, *15*, 31892. [[CrossRef](#)] [[PubMed](#)]
82. Belouadah, E.; Popescu, A.; Kanellos, I. A Comprehensive Study of Class Incremental Learning Algorithms for Visual Tasks. *Neural Netw.* **2021**, *135*, 38–54. [[CrossRef](#)] [[PubMed](#)]
83. Van de Ven, G.M.; Tuytelaars, T.; Tolias, A.S. Three Types of Incremental Learning. *Nat. Mach. Intell.* **2022**, *4*, 1185–1197. [[CrossRef](#)] [[PubMed](#)]
84. Ta, V.D.; Liu, C.M.; Nkabinde, G.W. Big Data Stream Computing in Healthcare Real-Time Analytics. In Proceedings of the 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, Chengdu, China, 5–7 July 2016; pp. 37–42.
85. Lomotey, R.K.; Pry, J.; Sriramoju, S. Wearable IoT Data Stream Traceability in a Distributed Health Information System. *Pervasive Mob. Comput.* **2017**, *40*, 692–707. [[CrossRef](#)]
86. Zvikhachevskaya, A.; Markarian, G.; Mihaylova, L. Quality of Service Consideration for the Wireless Telemedicine and E-Health Services. In Proceedings of the 2009 IEEE Wireless Communications and Networking Conference, Budapest, Hungary, 5–8 April 2009; pp. 1–6.
87. Van, D.D.; Ai, Q.; Liu, Q. Vertical Handover Algorithm for WBANs in Ubiquitous Healthcare with Quality of Service Guarantees. *Information* **2017**, *8*, 34. [[CrossRef](#)]
88. Chen, S.; Rehemat, T.; Li, C.; Wu, X.; Lin, H.; Lin, Z.; Liang, X.; Li, H.; Zhang, J. Hypertension Monitoring by a Real Time Management System for Patients in Community and Its Data Mining by Vector Autoregressive Model. *IEEE Access* **2023**, *11*, 12606–12621. [[CrossRef](#)]
89. Cartolano, A.; Cuzzocrea, A.; Pilato, G. Analyzing and Assessing Explainable AI Models for Smart Agriculture Environments. *Multimed. Tools Appl.* **2024**, *83*, 37225–37246. [[CrossRef](#)]
90. Lee, Y.W.; Choi, J.W.; Shin, E. Machine Learning Model for Diagnostic Method Prediction in Parasitic Disease using Clinical Information. *Expert Syst. Appl.* **2021**, *185*, 115658. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.