Ontology-based Data Preparation in Healthcare: the Case of the AMD-STITCH Project

Federico Croce[1†], Riccardo Valentini[1†], Marianna Maranghi[1], Giorgio Grani[1], Maurizio Lenzerini[1], Riccardo Rosati[1]

[1*]Sapienza Information-Based Technology InnovaTion Center for Health (STITCH), Sapienza University of Rome, Italy.

Contributing authors: federico.croce@uniroma1.it; riccardo.valentini@uniroma1.it; marianna.maranghi@uniroma1.com; giorgio.grani@uniroma1.it; maurizio.lenzerini@uniroma1.it; riccardo.rosati@uniroma1.it;

[†]These authors contributed equally to this work.

Abstract

In the context of healthcare, an AI solution is generally developed for a specific analysis task, based on a relevant dataset, with little attention to reusability and generalizability of its data preparation step. This paper focuses on a different scenario, which can be called *context-oriented*, where a set of clinical data sources, relevant for a specific context (e.g., a particular disease), is available and can be used for a variety of data analytics tasks, often carried out by different research groups. Therefore, the aim of this research is to present a systematic method, which exploits the Ontology-based Data Management paradigm to enhance data preparation in a context-oriented scenario. The introduced methodology has been applied to a project dealing with big data and regarding the treatment of diabetes and its complications. The peculiarity and challenge of this project lies in the fact that it deals with real world data, extracted from Electronic Medical Records within a 13 years timeframe, and thus not collected for research purposes. The paper focuses on two main steps of data preparation, namely **data modeling** and **data cleaning**, and it shows how this approach provides effective techniques for setting up a unified and shared database, to be used in the subsequent data analytics phases as an asset.

1 Introduction

Artificial Intelligence (AI) is increasingly prevalent in healthcare, due to the potential it has to transform many aspects of patient care, as well as administrative processes within healthcare organizations. However, it is well-known that Artificial Intelligence, and in particular Machine Learning, is not effective enough without a proper data preparation [16]. Data preparation is the process of gathering, transforming and cleaning raw data prior to processing and analysis. It is an important step in any data engineering and data science project, involving tasks such as understanding, collecting, reformatting, aggregating, integrating, combining and enriching raw source data and making modifications and corrections in order to meet quality standards in the target information system [14].

In the context of healthcare, an AI solution is generally task-dependent, i.e. developed for a specific analysis task, based on a particular dataset, composed by those features that are relevant for that task. In this "*task-oriented*" scenario, data preparation is often carried out through ad-hoc methods, with little attention to **reusability** and **generalizability**. As a result, the developed AI models are generally tightly coupled with the specific data source, file, or object store used to construct the dataset. In addition, the scenario is different also because of the type of collected data, being extracted from Electronic Medical Records (EMR) over a period of time of 13 years, and thus not collected for research purposes. This aspect forced to tackle several challenges, such as missing information, different interpretation of data, error detection, etc., paving the way towards adapting these strategies to other real-world medical contexts.

The present work deals with a different scenario, that can be called "*context-oriented*", where a set of clinical data sources that are relevant for a specific context (e.g., a particular disease) is available and can be used for a variety of data analysis tasks, often carried out by different research groups focusing on different **sub-projects**, within the same project. In this scenario, the task-oriented approach, characterized by a siloed nature, is ineffective. Conversely, a more solid and disciplined data preparation methods are needed, in order to foster:

- interoperability;
- common understanding of the semantics of data;
- coherence between different views of such data needed in the various sub-projects;
- accuracy and reliability of data in the integrated information system;
- compliance with data privacy requirements of the integrated information system.

We point out that all these features are coherent with the principles of the recent approach named "**Data-centric AI**" [3].

Therefore, the goal of this paper is **to present a method for data preparation in such a context-oriented scenario** and discuss its application in a project dealing with data related to the treatment of diabetes and its complications.

The followed approach is based on the **Ontology-based Data Management (OBDM)** paradigm [8, 19, 25], whose key idea is to manage the relevant data in a

particular context by resorting to a three-level architecture, constituted by the ontology, the data sources, and the mapping between the two. An *ontology* is a conceptual, formal description of the domain of interest in the given context, expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The *data sources* are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The *mapping* is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to enable a vast range of information services provided to users, and to express such services based on the conceptual description of the domain represented by the ontology. Exploiting this approach, the integrated view that the system provides to users is not merely a data structure accommodating the various data at the sources, but a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships governing them. The distinction between the ontology and the data sources reflects the separation between the **conceptual level**, the one presented to the user, and the **logical/physical level** of the information system, the one stored in the sources, with the mapping acting as the **reconciling structure** between the two levels. This separation brings several potential advantages.

Firstly, the ontology layer in the architecture is the obvious mean for pursuing a declarative approach to information integration, and, more generally, to data governance. By making the representation of the domain explicit, we gain re-usability of the acquired knowledge, which is not achieved when the global schema is simply a unified description of the underlying data sources.

Secondly, the mapping layer explicitly specifies the relationships between the domain concepts on the one hand and the data sources on the other hand. Such a mapping is not only used for the operation of the information system, but also for documentation purposes. The importance of this aspect clearly emerges when looking at large organisations where the information about data is widespread into separate pieces of documentation that are often difficult to access and rarely conforming to common standards. The ontology and the corresponding mappings to the data sources provide a common ground for the documentation of all the data in the organisation, with obvious advantages for the governance and the management of the information system.

A third advantage has to do with the extensibility of the system. One criticism that is often raised to data integration is that it requires merging and integrating the source data in advance, and this merging process can be very costly. However, the ontology-based approach we advocate does not impose to fully integrate the data sources at once. Rather, after building even a rough skeleton of the domain model, one can incrementally add new data sources or new elements therein, when they become available, or when needed, thus amortising the cost of integration. Therefore, the overall design can be regarded as the incremental process of understanding and representing the domain, the available data sources, and the relationships between them.

As we said before, the application of the OBDM approach to a project related to diabetes , one of the most common chronic diseases affecting hundreds of millions of people worldwide, is illustrated. This study takes advantage of one of the largest worldwide-available collections of diabetic patient records, the Associazione Medici Diabetologi (AMD) dataset, obtained from electronic medical records of Italian diabetes patients. Such a dataset was recently made available by the AMD and the AMD Foundation to the Sapienza information-based Technology InnovaTion Center for Health (STITCH) at Sapienza University of Rome.

The data preparation phase was carried out using the OBDM approach, whose result is a structured, cleaned database with all the data provided by AMD, and available for different types of data analytics tasks. The main information services that have been employed in this project are related to the modeling of a unified database and the cleansing of the data. Both services heavily relied on the ontology defined following the OBDM paradigm. In turn, the definition of the ontology and the modeling of meta-data have been carried out with an extensive interaction with physicians and domain experts. The application of this methodology for the design of the AMD data layer can be set as a gold standard to deal with the healthcare big data, as they are domains in which ontologies and a thorough data preparation can dramatically enhance data quality and improve the reliability of the subsequent analyses.

The paper is organized as follows. Section 2 is a survey on recent research in data preparation in healthcare, focusing on the approaches that deal with data obtained from Electronic Health Records. Section 3 goes into the details about AMD and the process used to gather the data provided to the project. Sections 4 and Section 5 describe the structure of the raw data used as input to our methodology, thus illustrating the two main steps in the data preparation methodology, namely data modeling and data cleaning. Section 6 is a discussion on how future data analytics tasks can take advantage from the data preparation step, and it provides an evidence of the importance of our data preparation method, in order to carry out high quality data analytics. Section 7 concludes the paper by mentioning future research directions within this project.

2 Related work

Data preparation is a crucial step in healthcare data analysis, as it ensures that the data is accurate, complete, and ready for analysis. Some of the most important aspects of data preparation in healthcare can be summarized as follows:

- **Data cleaning**. This involves identifying and correcting errors in the data, such as missing values, incorrect data types, and outliers. Cleaning the data is essential to ensure that it is accurate and complete before analysis.
- **Data integration**. Healthcare data often comes from multiple sources, such as electronic health records, claims data, and patient surveys. Integrating this data into a single, unified dataset is important for analysis, and it can also help to identify patterns and relationships which would be missed otherwise.

- **Data transformation**. Transforming data involves converting it into a format which is suitable for analysis. This may include aggregating data, converting data types, and normalizing data.
- **Data reduction**. Reducing the data involves identifying and removing irrelevant or redundant data. This can help to improve the performance of analysis algorithms and to reduce the risk of overfitting.
- **Data anonymization**. Healthcare data contain sensitive information that must be protected. Anonymizing the data involves removing or masking identifying information, such as names and addresses, to protect patient privacy.
- **Data validation**. Validation involves checking the data for errors or inconsistencies, such as data falling outside expected ranges or data violating some business rules. Validating the data can help to ensure that they are accurate and reliable for analysis.

Overall, these approaches are essential for ensuring that healthcare data are accurate, complete, and ready for analysis, which can ultimately improve patient care and outcomes.

In the last years, many papers have dealt with the specific problem of EHR data preparation, i.e. the problem of preparing datasets directly obtained from Electronic Health Records (EHR) and Electronic Medical Records (EMR).

One of the earliest works dealing with the problems related to data preparation in the context of EHR data is [20], which explicitly states that preprocessing and transformation of data are necessary preconditions for data analysis (in this case, data mining) of clinical data. The paper proposes an approach to data preparation, which uses information from data, metadata and sources of medical knowledge.

Then, [27] surveys the application of deep learning to clinical tasks based on EHR data, identifying several limitations of current research with respect to models' interpretability, data heterogeneity, and availability of universal benchmarks.

Given the high variability of the quality of EHR data, [26] proposes a new framework to evaluate the suitability of clinical dataset to satisfy the research needs of observational studies.

Moreover, [11] proposes a new approach to the integration of genetic data and EHR data. In particular, it proposes a novel method to scan phenomic data for genetic associations using International Classification of Disease (ICD-9) billing codes.

A very recent contribution is [23], which presents a general approach to data preparation for the analysis of EHR data. In particular, it starts from two observations: (i) the absence of a framework which integrates data quality assessment and cleaning methodologies in an iterative workflow; (ii) the lack of schemes and representative case studies for assessing and validating cleaning methods. The paper thus proposes a new EHR data preparation framework to guide EHR data quality assessment and cleaning workflows. Such an approach solves some of the problems related to the assessment and validation of cleaning methods for EHR data.

Finally, [29] focuses on the need for systematic methods for EHR data quality assessment, and proposes a new dynamic, evidence-based guideline to enable EHR data quality assessment, called 3x3 DQA, based on the three primary dimensions of

EHR data (patients, variables, and time), three main aspects of data quality (complete, correct, and current data), and on the idea that data quality is task-dependent.

On a different level, [17] proposes a novel information architecture for healthcare information systems, with the purpose of decreasing semantic ambiguity of data.

As for previous work on ontologies for the diabetes context, we refer to Section 4, where we analyze, and compare with our approach, some of the most relevant works in this area, in particular [12] and [13].

3 The AMD dataset and the AMD-STITCH research goals

In this section we provide a description of the process used to gather the data provided to our project.

People with diabetes have an increased risk of developing life-threatening health problems (micro and macro vascular complications), which results in reduced quality of life, increased mortality and higher healthcare costs [5, 21]. Type 2 Diabetes (T2DM) is the most common type, accounting for around 90% of all diabetes cases [4]. The burden of T2DM is increasing worldwide at earlier ages, due to the widespread distribution of predisposing factors such as obesity and sedentary lifestyle [10, 21]. Moreover, due to the impact that pandemics are having on the general economy and on the investments that will be made in Health, there is an urgent need to better allocate available economic resources. As an example, in the case of T2DM, it is necessary to improve medium and long term risk prediction systems, to identify subgroups of subjects which will not develop diabetic complications, assess the potential of currently used drug therapies to modify final outcomes, or identify subgroups of subjects that are not responding to these expensive drugs [24].

The policy for the management of subjects with diabetes in Italy has its own uniqueness if compared to the rest of the world. The *Associazione Medici Diabetologi* (AMD) is the largest Italian professional national society serving clinicians who are challenged with the treatment of diabetes and its complications. Since 2005, AMD has promoted the creation of a network of diabetes outpatient clinics sharing the same electronic medical record used for the management of patients in order to promote and improve the quality of care of diabetic subjects. Periodic data extraction from the EMR has been used for monitoring the quality of care indicators [6]. This activity has produced over the years a improvement of all the process indicators considered and has proved to be cost-effective [6, 9]. Today the network comprises about 250 diabetic centers and results are published every year in the AMD annals [6].

The AMD dataset represents a valuable source of observational research data allowing to deepen our knowledge of many key aspects of this chronic disease. A recent agreement between AMD and Sapienza information-based Technology InnovaTion Center for Health (STITCH) has made available at Sapienza a dataset containing the data of about 600,000 patients collected over a time interval of 13 years (2005-2018).

More in detail, the EMR is used for the daily management of patients in the outpatient clinics: data that may concern different specifications of the patient taken in charge by each Center (for example demographics, laboratory exams, diagnostic tests

for the screening of the complications, drug prescriptions) are entered periodically (at least once a year), by various professional figures in the same medical records (doctors, nurses, dietitians). This, with the aim of recording the clinical progress of the disease, therapies, and specific tests ordered for each patient. Over the years, the EMR has been modified both in its design (more and more user-friendly), and in its content, with the possibility of making some fields mandatory and classifying some features as permanent (e.g., the presence of a complication that changes the patient's clinical status). It should be also noted that, due to the EMR structure, some of these data could be historicized (e.g., the trends of some patient parameters), while other data cannot (e.g., the working status of a subject from active to retired).

Moreover, in the years in which AMD data has been collected, the medical attitude towards diabetes management has been changing due to new scientific evidences. This implies that the type and the variety of the data collected in the AMD dataset is unique: it is based on current practices and it represents real world data. For this reason, using those data for research purposes represents also a challenge: information is fragmented and it does not constitute a dataset built for scientific purposes, as it reflects a clinical and data recording logic that has changed over the years.

With these premises, by leveraging machine learning approaches in order to possibly implement the treatment of this chronic disease and reducing the costs, the identified research aims are:

1. study the potential impact on micro and macro vascular disease (CVD) outcomes of time-dependent covariants, including new parameters like weight gain in predicting or modifying the outcomes;
2. assess the potential of currently used drug therapies to modify CVD outcomes over time and identify subgroups of subjects that are not responding to these expensive drugs.

Due to the nature and complexity of the data contained in AMD dataset, in each step that will be described below, close and continuous collaboration was required between physicians, who are experts in the EHR usage and in the specific domain, and computer science engineers. It is important to highlight that, without this continuous knowledge sharing, the systematization and understanding of the data would not have been possible.

4 Data Modeling

The data modeling activities of the project are described in this section of the paper. The main goal of these activities is the definition of the schema of the database representing the shared common knowledge, to be used in the data analytics tasks. Before delving into the details of our techniques, the motivations that drove this research into such a thorough reorganization are illustrated.

4.1 The initial dataset

The data from AMD came in the form of several *CSV* files and one *PDF* document describing them. Even though the PDF document was meant to describe the content of the CSV files, the data resulted extremely difficult to inspect and interpret.

The CSV files were organized in such a way that even very different concepts were in fact stored in the same file, and only distinguished by the value associated with an attribute, whose meaning had to be searched, mostly manually, in the PDF file. For example, both the prescriptions of drugs, and disease diagnoses were stored in the same file. To distinguish between these two cases, one had to look at the values associated with a specific attribute contained in the CSV file. If the associated value was the Anatomical Therapeutic Chemical code [2] of a drug, it indicates that the corresponding row in the CSV file was referring to the prescription of a drug. Otherwise, if the corresponding value was the character "S", the corresponding row in the CSV file was referring to a disease diagnosis.

Clearly, this way of interpreting the meaning of the different rows in the CSV files was confusing and little informative, thus motivating the work presented in this paper.

4.2 Ontology-based schema design

For the design of the database schema, the *Ontology-based Data Management (OBDM)* approach [25] was followed. The idea of OBDM is to manage a set of data sources through an *ontology*, i.e. a shared, virtual conceptualization of the domain of interest of the data sources, and through declarative *mappings* that link the data sources to the ontology.

In our project, the OBDM approach has been exploited to generate a new database schema to properly represent the AMD dataset. More precisely, the following steps have been carried out:

1. *Domain and dataset analysis* - the domain of interest and the AMD dataset was analyzed, with the goal of understanding and specifying the semantics of the dataset.
2. *Meta-data analysis* - A large collection of meta-data have been crucial for the understanding of the AMD dataset. Indeed, such a dataset cannot be understood without the information about the medical standards (e.g., ICD-9-CM encoding, proprietary AMD encoding, etc.), used for encoding data in the original dataset. This information is actually not materialized in the AMD dataset: it was mostly available only through PDF files. To improve the quality of the new representation of the data, the decision has been to explicitly store a suitable representation of such meta-data in the database.
3. *Ontology design* - an ontology (in the OWL language [7]) has been defined, modeling both the data and the meta-data of the domain of interest.
4. *Database schema design* - Based on the ontology built in the previous step, a database schema has been defined; this schema is able to consistently represent both data and metadata.

It has to be pointed out that all the above phases have required a tight cooperation between domain experts and data engineers. In particular, given the complexity of the domain of interest and the relatively low level of abstraction of representation in the dataset, the domain and dataset analysis has been a very challenging task, requiring not only the involvement of several physicians, but also the participation of experts of the dataset itself. In addition, the definition and validation of the domain ontology has been an iterative and interactive process, heavily involving the stakeholders.

4.3 The ontology

The ontology has been formalised using the W3C Web Ontology Language (OWL) [7, 8] and it consists of all the major relevant concepts and relations together with all the characterizing properties. Figure 1 represents a snippet of such an ontology.

The figure uses "Graphol", a graphic formalism that allows one to view the OWL ontologies in a diagram [18]. These are the conventions adopted:

- The concepts and attributes in red are those for which there is not any information about instances. For example, there are no known properties of the AMD centers such as address, province, email, etc., or of the people you do not know properties such as name, surname, responsibilities of AMD centers, and so on.
- The concepts colored in green are concepts whose instances have rigid properties (i.e., properties that do not change over time, such as for example a diagnosis of diabetes for a Patient) or not rigid (i.e., properties that can change over time, such as for example the marital status of a patient) but that are not historicized.
- Brown-colored concepts are historicized, i.e. their instances provide a history of a particular phenomenon.
- The concepts colored in orange describe the metadata. In Figure 1 some example instances of the metadata concept "ExaminationType" which represents all the different types of examination that are collected in the dataset, are provided. In the example depicted, seven different types of examination are shown, namely the asymptomatic hypoglycemia, the symptomatic hypoglycemia, the foot and GISED questionnaires, the specific and overall medical history, and the severe hypoglycemia types. Each of these types are related to specific sub concepts of the more general "Examination" concept, that represent the actual examination instances conducted on specific dates and with associated examination values. By linking each specific sub concept with a corresponding instance of the metadata concept "ExaminationType", the fact that all examination instances of the same type have the same associated relevant metadata attributes such as the "AMD code", and the "ICD9-CM code", is modeled.
- Non-colored concepts are those that do not have direct instances (their instances are those of sub-concepts or are derivable by rules).

The ontology shown in the diagram, although incomplete (for example, not all metadata are described in the ontology) clarifies the context in which the data of the database must be interpreted. Here, a brief description of this context is provided.

The data come from the systems used by the doctors of the various centers to collect information on the patients managed by the center. The concept at the center

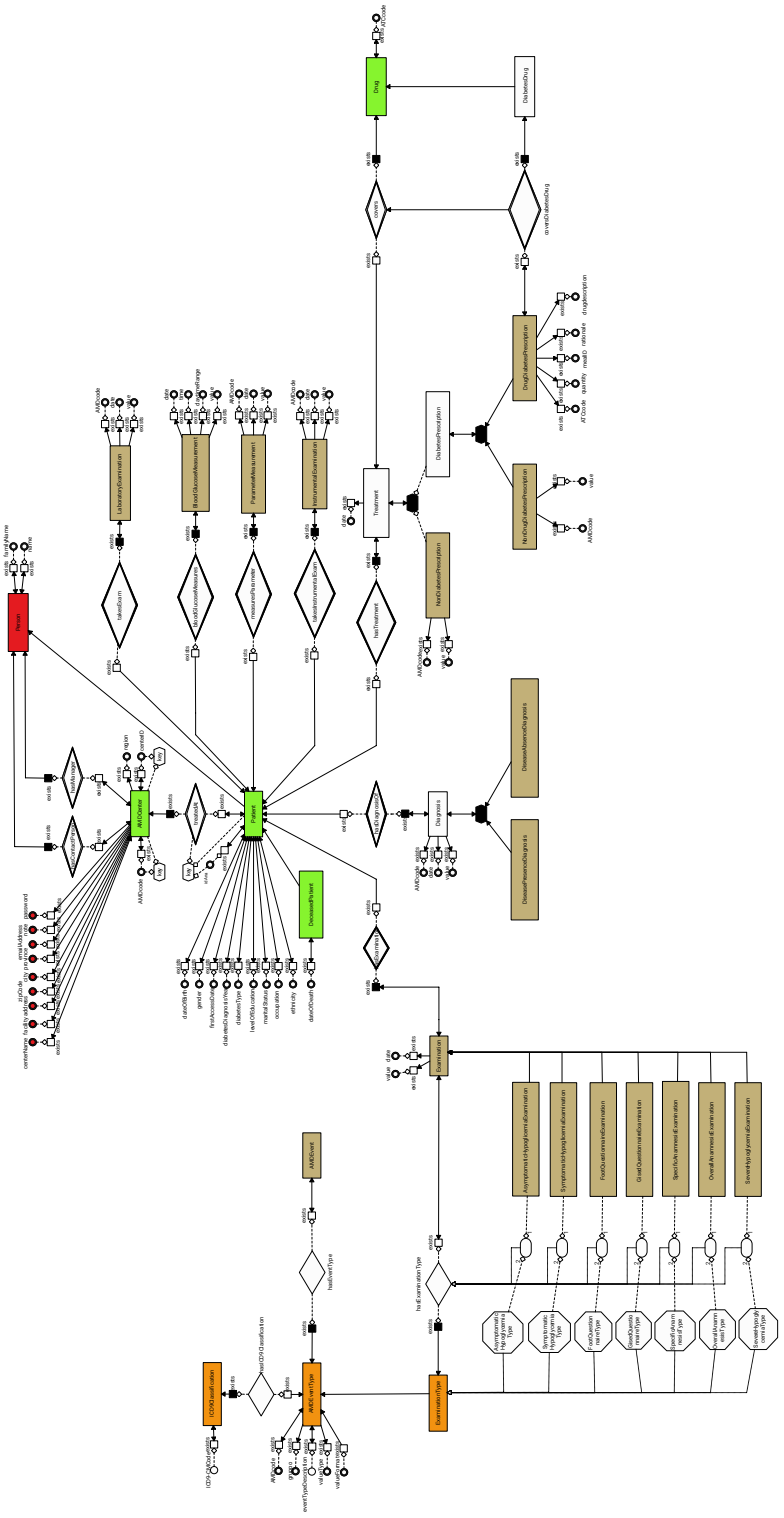

Fig. 1 The ontology representing the main concepts and relations involved in the data published yearly by AMD, as well as an example of the metadata describing them.

of the domain is therefore "`Patient`". Each patient (instance of the Patient concept) belongs to an AMD center and has various properties represented by attributes. "`PatientDeceased`" is obviously a sub-concept of Patient and has "`date of death`" as an additional property. Patients are the protagonists of a series of events, whose historical succession is represented by the instances of corresponding concepts.

- Patients are subject to visits (concept `Visit`).
- Patients are subject to diagnosis (`Diagnosis` concept and related sub-concepts).
- Patients perform different types of exams (concepts `Laboratory Exam`, `Blood Glucose Measurement`, `Parameter Measurement`, `Instrumental Exam`).
- Patients are assigned prescriptions by doctors (`Prescription` concept and its sub-concepts).
- Prescriptions can be related to drugs (`Drug` concept and related sub-concepts).

Despite the existence of DMTO ontology that models the *diabetes mellitus domain* as a whole [12] [13], the adoption of a novel ontology is justified by the presence of specific characteristics entailed by the AMD domain. For example, each instance of the AMD ontology concept "`Laboratory Exam`" has a corresponding "*AMD code*" which should be taken into account, in this domain description, as a data property. The DMTO ontology has the general purpose of being a comprehensive knowledge base for a theoretical description of the diabetes domain, as the main manifested goal of DMO and DMTO is to incorporate the knowledge concerning the diabetes disease in order to enhance a *clinical decision support system*. Therefore, DMTO can be seen as a top-level ontology to provide clinicians with a powerful tool helping them with diagnostics procedures and treatment plans for their patients. On the one hand, it can be considered a gold standard for the *description of the disease*, as it contains complications, laboratory tests, symptoms, physical exams, demographics, diagnoses and treatment; on the other hand, in the case of the AMD domain, it needs an unavoidable extension to introduce knowledge describing other notions, e.g. the ones of AMD Centers and ICD-9 codes related to diagnoses and procedures.

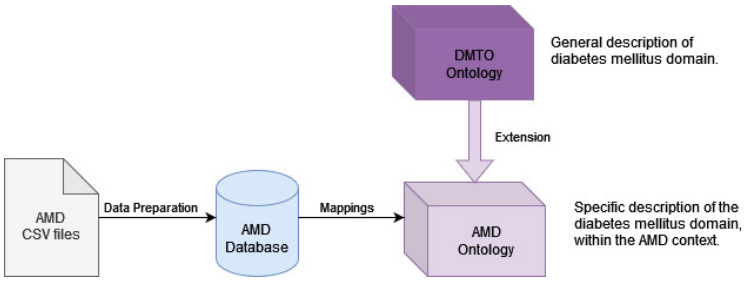

Fig. 2 The role of the AMD ontology: between the real-world data contained in the AMD database and the general conceptual description of the diabetes disease provided by DMTO.

The AMD ontology stands in the middle: it provides an effective modeling of the AMD electronic record data and metadata and it could allow, through specific

mappings, to frame those real-life data within the more general descriptions of the DMTO ontology. In this way, it is possible to reach two objectives:

1. *providing an effective and immediate description of the AMD domain, accomplishing all the context-specific prerequisites (AMD perspective);*
2. *framing this description in a more disease-related domain, possibly enhancing interoperability with other T2DM research projects (external perspective).*

Finally, it should be noted that the overall ontology (of which the diagram shows the fundamental core) contains all the sufficient and necessary elements for a possible construction of a Knowledge Graph that expresses all the knowledge (intensional and extensional) encoded in the data and the metadata.

4.4 The database schema

Starting from the formalization of the domain represented by the ontology, a database schema was designed, trying to stay as close as possible to the conceptualization represented by the ontology. More specifically, the database schema is composed by two parts: data and meta-data. The former part of the schema (data tables) was meant to contain the actual data, whereas the latter (meta-data tables) contains meta-data information such as the type of values that are admissible for each kind of data, the range of such values, the national level code associated with each kind of medical examination (if available), and so on.

4.4.1 Data tables

The data tables are the following:

1. `data.amdcenters`: registry of the centers;
2. `data.registry`: patient registry;
3. `data.visit`: records "visit" events;
4. `data.laboratoryparameterexam`: records laboratory examination or parameter detection events;
5. `data.glycemia`: records capillary glucose measurements performed by the patient at home;
6. `data.instrumentalexam`: records instrumental examination events;
7. `data.diagnosis`: records "diagnosis" events, which in addition to diagnoses related to diabetes, also include comorbidities and complications;
8. `data.diabetesdrugprescription`: records events of prescription of drugs as part of the treatment of diabetes;
9. `data.nondrugdiabetesprescription`: records prescriptions related to blood glucose control and non-drug related prescriptions (for example, the diet);
10. `data.nondiabetesprescription`: records prescription events that do not concern the treatment of blood sugar, but the comorbidities most frequently associated with diabetes (for example treatment of hypertension and high cholesterol levels).

4.4.2 Meta-data tables

The meta-data tables are one for each data table. The meta-data tables for the data tables 1, 2, 5 and 8 have a row for each field of the data table of the same name, and this row shows the information on the corresponding field. The other meta-data tables, on the other hand, report information on the possible types of records contained in the data table of the same name, in relation to the possible values of the codiceamd field.

ICD-9-CM codes appear in some of the above tables. ICD-9-CM relates to the International Classification of Diseases (ICD), a classification system that organizes diseases and injuries into groups on the basis of defined criteria. [22] We point out that, even though updated versions of the ICD9-CM classification exists, in fact the 11th release of the classification system has been updated on January 2022, in this project we stick with the 9th release because it is the classification system used in the AMD dataset at the moment. Our framework is insensible to the classification version used, and will eventually be updated to more recent versions as soon as the AMD dataset will also be updated.

5 Data cleaning

In this section, the cleaning process carried out on the AMD dataset is described[15]. It is important to note that, in this project, data cleaning is intended as a continuously improving process that collects feedback from the data analytics tasks, and use them to improve the quality of the dataset. For example, the actions described in section 5.2 have only been made necessary after a data analytics task used the data resulting from a first iteration of our cleaning process.

The original data received from AMD contained several problems, inconsistencies and errors. These poor data quality problems are mainly due to the fact that AMD gathered them from real medical examinations coming from almost 300 different centres for diabetes treatment in Italy since 2005[1]. Along these years, different versions of the EMR software tool have been used for collecting data, and even the same version of such a tool has been used in different ways in the various centers: this caused several semantic discrepancies within the gathered data.

Many cleaning actions were necessary to overcome these problems. The actions taken can be distinguished in three main areas: actions to resolve errors in the data, actions to resolve missing and incomplete data, actions to preserve data privacy. For all the actions, the main criteria that have been adopted were to keep as much information as possible (or equivalently to delete the least amount of data), and to consider only the quality checks that could be done using one single table at a time, thus postponing cross-table checking to the future work. The first criterion can be considered as a natural approach to all data cleaning tasks, and the decision to adopt the second criterion is motivated by the fact that since at this stage the aim was to keep each table consistent with itself and with its metadata descriptions.

[1] We clarify that not all the 320 centres constituting the AMD network included their data in the latest AMD annals.

5.1 Actions to resolve errors in the data

An important objective of the cleaning was to make data and meta-data descriptions consistent with each other. Whenever a discrepancy between what was described in the meta-data the stored data occurred, two different strategies could be adopted: changing the data to make them consistent with their meta-data description, or changing the meta-data to effectively describe the actual content of the data. It has not been possible (neither it was reasonable), to always adopt one of the two strategies. As a result, this kind of error has been analyzed on a case-by-case basis, with a strong collaboration with the stakeholders, i.e. the physicians of our group.

For example, the meta-data has been considered correct in stating that the patient's year of birth should always be later than 1900, and all data regarding patient's year of birth that were prior to 1900 have been set to NULL. On the other hand, the meta-data for some laboratory exams reported the wrong number of allowed digits in the corresponding values. These cases have been identified and, after establishing that the data were correctly reporting values with the expected number of digits, the number of allowed digits in the meta-data for such laboratory exams have been corrected.

5.2 Actions to resolve missing and incomplete data

As one might expect from a dataset of this magnitude and complexity, the data reported many missing values in the form of NULLs. With the goal of improving the overall quality of the dataset, there was an attempt to resolve these problems by populating the values whenever possible. For example, some data in the table "`data.diagnosis`" represents a disease diagnosis made by a physician. There are two types of diagnoses of this kind in the dataset, either positive, meaning that the physician found the presence of the disease, or negative, meaning that the physician explicitly verified the absence of a disease. Since in both cases, the mere presence of the tuple manifested a clear intention of the physician to express a disease diagnosis (either positive or negative), the actual value associated with these tuples were irrelevant. Nevertheless, very different cases were found for these types of tuples:

- positive disease diagnosis with associated a NULL value;
- positive disease diagnosis with associated a value representing the ICD-9-CM [22] code of the corresponding disease;
- positive disease diagnosis with associated the value "S";
- negative disease diagnosis with associated a NULL value;
- negative disease diagnosis with associated a value representing the ICD-9-CM code of the corresponding disease;
- negative disease diagnosis with associated the value "S".

For the above-mentioned reasons, all the listed cases have to be considered equivalently, although knowing the specific ICD-9-CM code [22] of a disease diagnosis is obviously an added value. Therefore, all NULL values were removed from the table "`data.diagnosis`" associated with these types of tuples, by substituting them with either the ICD-9-CM code of the corresponding diseases, when possible, or simply the value "S" otherwise.

5.3 Actions to preserve data privacy

Ensuring the privacy of the patients in the AMD dataset turned out to be a non trivial task. In fact, even though all personal information of each individual had already been removed from the dataset, in many niche cases, it was still possible to trace the identity of people, given that one knew easily accessible information of them such as the region of the center they are taken care in. Privacy was considered as a central property in this dataset, therefore several actions were taken in order to mitigate the risks of identity disclosure:

- Removal of the specific day and month from the date of birth of patients, by only leaving the more general year of birth.
- Removal of the specific day and month from the date of first access to a center in the AMD network.
- Removal of the specific day and month from the date of first diagnosis of diabetes.
- Substitution of the code of the associated diabetes center with a new fictitious value, for all cases in which it was possible to identify the specific person by joining together other known information such as her gender, year of birth, and the region of the diabetes center she is being treated. By doing this substitution, all patients that were at risk of a privacy leakage into one single bigger fictitious center were collected. As a result, it has been verified that all patients that were recognizable in their original center are now indistinguishable from other patients in the new center.

6 Evaluation and discussion

In this section, a brief introduction about the role of the data analytics within AMD domain context is provided, examining the currently involved research tasks, to then describe an example of evaluation of a simple data analytics task, on two settings: a relational database directly derived from the data sources provided by AMD and the relational database coming from the data preparation discussed before.

6.1 The role of the data analytics on AMD data

One of the peculiarities of the AMD dataset, which reinforces the need for the whole data preparation process, is the fact that the involved data have to be accessed in order to accomplish heterogeneous data analytics tasks, each one with specific research targets. Therefore, ideally, data have to be *conceptualized* and *globally cleansed* only once, allowing any kind of analysis to look at the very same set of information. Orthogonally, the refinement of data can derive by the discovery of novel inconsistencies, acknowledged during the exploration of the dataset.

In what follows, a list of *five* different data analytics tasks is presented. They are managed by different research groups, but they all focus on T2DM, the most prevalent form of diabetes.

1. Data-driven assessment of treatment response and disease progression, to identify potentially clinically important differences among patients;

	csv files db	final db
Number of records	1,413,261	1,174,909
Average HbA1c (%)	121,394,886.00	7.48

Fig. 3 Query results over the initial dataset and over the database obtained by our approach.

2. Mining recurrent patterns in patients' histories for diabetes progression mode, to define a disease progression model (DPM) for diabetes, and to use this progression model to define groups of patients having a common disease progression, and thus a common response to treatment;
3. Onset short term prediction of retinopathy and nephropathy in order to act at the level of primary prevention;
4. Patient clustering based on trajectory of glycated hemoglobin (HbA1c), to investigate its trajectories along the years, with the aim of understanding the effects of treatments;
5. Short-term prediction of cardiovascular events such as myocardial infarction, coronary angioplasty, and others.

6.2 Example of comparison between the data before and after the data preparation

Glycated hemoglobin (HbA1c) is the test of choice for diagnosing diabetes and monitoring glycemic control: raised HbA1c levels are associated with micro and macrovascular diabetic complications. In order to understand the effects of treatments on a patient, it is crucial to investigate her trajectories of HbA1c values during a certain time frame.
To give an example of the enhancement coming from the whole process of data preparation discussed so far, one has to imagine the need *to return all the demographics of the patients, coupled with their sequences of HbA1c values along the years* (this involves only a small subset of the AMD dataset).

In order to evaluate the differences between performing this task in a database directly derived from the data sources and in the stable version of the AMD database, i.e. the one obtained from the ontology conceptualization of the domain and the data cleaning described so far, a relational database in which each csv file resulted in a table was set up. In this kind of scenario "*all the personal data of the patients*" are split into three different tables of anagraphic data and "*their sequences of HbA1c along the years*" is an aggregation query, also merging data split into other three different tables of numeric information. On the other hand, in the setting of the conceptualized and clean database, there exists only one table hosting the patients' demographics and another one containing the HbA1c measures required by the task.

Besides the qualitative analysis of the query complexity, a further step of comparison follows the execution of both queries (having the same semantics) and consisted of investigating the presence of relevant differences between the two settings. Those data are reported in the table of Figure 3.

This example of task showed some important differences between the setting of the *source-based database* and the *final database* coming from the data preparation discussed in this paper.

Comparing the number of records, it resulted that a **16.8%** of them is missing in the final database. This is a direct consequence of the data cleaning, which caused the removal of duplicates or totally incorrect records in some cases. It means that a significant portion of the records returned in the setting of the source-based database contained some sort of inconsistency.

The most interesting result obtained is represented by the second row of the table. In fact, a very elementary analysis was imagined to be performed on the sequences of the HbA1c: *returning the mean of the sequence for each patient (a record) and to compute the average of these mean values*. In the case of the final database tables, the outcome (7.48%) is plausible according to the clinical literature guidelines about HbA1c; vice-versa, in the case of the csv files, the outcome (121,394,866%) is totally inconsistent [1]. Performing a more in-depth analysis of the *source-based database setting*, it resulted that many records were characterized by both plausible HbA1c values and totally *out of scale*, and thus inconsistent, ones.

These observations suggest that querying the csv data sources without any data preparation produces a result with potentially several inconsistencies that, to be overcome, require a tight collaboration with the domain experts. In other words, in the case of holding a database directly derived from the *csv data sources*, each analysis should require a proper data preparation step (assisted by the stakeholders) before obtaining similar results w.r.t. the ones coming from querying the current *final database*. The difference between this ad hoc data preparation process and the one discussed in this paper is that the first one is indeed *task-dependent*, while the latter is general as it comprises the database as a whole and, therefore, it is *context-oriented*.

7 Conclusions

In this paper, the OBDM approach has been applied to the AMD-STITCH project by organizing, cleaning, and making the AMD dataset the largest shared asset, made of real-world data, for all upcoming data analytics tasks related to diabetes research.

Many iterations of cleaning process were required before settling into a stable state of the database. It is clear that many more processing actions might be necessary in the future, both driven by the specific needs of upcoming data analytics tasks based on the data, and by future corrective maintenance that are inevitable with datasets of this size. Besides, AMD might release in the future an updated version of its dataset, thus requiring further updates, including part of the data preparation already presented, and its possible extensions.

As already pointed out, five different tasks, based on the refined AMD dataset, are being currently in progress. Although this data preparation work is encouraging, the real challenge would be to establish its actual effectiveness in supporting data analytics. This is the most important future direction of the present research: *if clinically relevant results were obtained through the exploitation of the AMD database, then the effectiveness of this approach would be confirmed*. Towards this goal, it would also be

interesting to apply the methodology presented in [23] for data quality assessment (See Section 2) to the outcomes of our approach. The initial attempts in this direction have shown that, to reach the above goal, further work is needed to adapt and integrate the two approaches. From the point of view of the ontology, as previously mentioned, it could be useful in the future to provide a mapping of some entities and relationships to the ones modeled by DMTO in order to further enhance interoperability.

As a final important remark, despite some peculiarities of the AMD case, many other application scenarios in healthcare might have similarities with the AMD project (see e.g. [28]), and therefore they can take advantage from following the presented methodology.

Acknowledgments

This work has been partially supported by MUR under the PRIN 2017 project "HOPE" (prot. 2017MMJJRE), by the EU under the H2020-EU.2.1.1 project TAILOR, grant id. 952215, and by the projects FAIR (PE0000013) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

The authors would like to thank the Associazione Medici Diabetologi (AMD), Fondazione AMD and all the scientists involved in the STITCH-AMD initiative for supporting this work. This work would not have been possible without the precious efforts of Dr. Sebastiano Filetti and the expertise of Dr. Antonio Nicolucci and Dr. Giuseppe Lucisano (CORESEARCH S.r.l.) and all the patients who have been cared over the years in the AMD centers.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] ADA - understanding A1C. https://diabetes.org/diabetes/a1c. Accessed: 2022-08-21.

[2] ATC code. https://www.ema.europa.eu/en/glossary/atc-code. Accessed: 2022-08-21.

[3] Data-centric ai. https://datacentricai.org. Accessed: 2022-08-21.

[4] International Diabetes Federation - about diabetes. https://www.idf.org/aboutdiabetes/type-2-diabetes.html. Accessed: 2022-08-21.

[5] International Diabetes Federation - facts figures. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. Accessed: 2022-08-21.

[6] The journal of amd. https://www.jamd.it/archivio-annali-amd/. Accessed: 2022-08-21.

[7] OWL web ontology language guide. https://www.w3.org/TR/2004/REC-owl-guide-20040210/, 2004.

[8] D. Calvanese, G. D. Giacomo, D. Lembo, and et al. Ontologies and databases: The dl-lite approach. In *Reasoning Web. Semantic Technologies for Information Systems*, pages 255–356. Springer, 2009.

[9] D. Cucinotta, A. Nicolucci, A. Giandalia, and et al. Temporal trends in intensification of glucose-lowering therapy for type 2 diabetes in italy: Data from the amd annals initiative and their impact on clinical inertia. *Diabetes research and clinical practice*, 2021.

[10] D. Dabelea, E. J. Mayer-Davis, S. Saydah, and et al. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. *JAMA*, 311(17):1778–1786, May 2014.

[11] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinform.*, 26(9):1205–1210, 2010.

[12] S. El-Sappagh and F. Ali. Ddo: a diabetes mellitus diagnosis ontology. *Applied Informatics*, 3(5), 2016.

[13] S. El-Sappagh, D. Kwak, F. Ali, and K.-S. Kwak. Dmto: a realistic ontology for standard diabetes mellitus treatment. *Journal of Biomedical Semantics volume*, 9(8), 2018.

[14] T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. Paton. Data wrangling for big data: Challenges and opportunities. In *Advances in Database Technology — EDBT*, pages 473–478, 2016.

[15] F. Geerts, G. Mecca, P. Papotti, and D. Santoro. Cleaning data with llunatic. *VLDB J.*, 29(4):867–892, 2020.

[16] M. Hameed and F. Naumann. Data preparation: A survey of commercial tools. *SIGMOD Rec.*, 49(3):18–29, 2020.

[17] K. L. Hua Guo, Michael Scriney. An ostensive information architecture to enhance semantic interoperability for healthcare information systems. *Information Systems Frontiers*.

[18] D. Lembo, V. Santarelli, D. F. Savo, and G. D. Giacomo. Graphol: A graphical language for ontology modeling equivalent to OWL 2. *Future Internet*, 14(3):78, 2022.

[19] M. Lenzerini. Managing data through the lens of an ontology. *AI Magazine*, 39(2):65–74, 2018.

[20] J.-H. Lin and P. J. Haug. Data preparation framework for preprocessing clinical data in data mining. In *AMIA Annu Symp Proc.*, pages 489–493, 2006.

[21] X. Lin, Y. Xu, X. Pan, and et al. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Scientific Reports*, 10(1):14790, Sep 2020.

[22] Medicode. ICD-9-CM: International classification of diseases, 9th revision, clinical modification. 1996.

[23] Z. Miao, M. D. Sealey, S. R. Sathyanarayanan, D. Delen, L. Zhu, and S. Shepherd. A data preparation framework for cleaning electronic health records and assessing cleaning outcomes for secondary analysis. *Inf. Syst.*, 111:102130, 2023.

[24] B. Pintaudi, A. Scatena, G. Piscitelli, and et al. Clinical profiles and quality of care of subjects with type 2 diabetes according to their cardiovascular risk: an observational, retrospective study. *Cardiovascular Diabetology*, 20(1):59, Mar 2021.

[25] A. Poggi, D. Lembo, D. Calvanese, and et al. Linking data to ontologies. *J. Data Semant.*, 10:133–173, 2008.

[26] N. Shang, C. Weng, and G. Hripcsak. A conceptual framework for evaluating data suitability for observational studies. *J. Am. Medical Informatics Assoc.*, 25(3):248–258, 2018.

[27] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Informatics*, 22(5):1589–1604, 2018.

[28] R. Valentini, E. Carrani, M. Torre, and M. Lenzerini. Ontology-based data management in healthcare: The case of the italian arthroplasty registry. In R. Basili, D. Lembo, C. Limongelli, and A. Orlandini, editors, *AIxIA 2023 – Advances in Artificial Intelligence*, pages 88–101, Cham, 2023. Springer Nature Switzerland.

[29] N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng. A data quality assessment guideline for electronic health record data reuse. 5(1), 2017.