Chapter 4

Association of Italian Diabetologists scenario

The Association of Italian Diabetologists[1] (Associazione Medici Diabetologi, AMD) was established on July 13, 1974, in Montecatini by a group of qualified representatives of Italian clinical diabetology at that time. With over 2000 members, it is the largest scientific association in Italian diabetology. Affiliated with the International Diabetes Federation (IDF), AMD:

- promotes the spread of facilities across the territory suitable for the prevention, diagnosis, and treatment of diabetes mellitus;
- is involved in the professional qualification and cultural updating of healthcare personnel operating in such facilities;
- works to ensure that diabetology and the figure of the diabetologist acquire and maintain their autonomy from both an educational and clinical perspective, constituting the primary point of reference in the care of diabetic patients.

AMD encourages research in the field of diabetology, both clinical and therapeutic, and collaborates with other institutions that share common goals and interests.

Analyzing the themes of various AMD Congresses, it becomes clear that since the 1990s, topics that are now "consolidated" have emerged as subjects for in-depth exploration and debate. These include themes such as "*diabetes and the quality of care*", "*the diabetologist and the family doctor in the integrated management of diabetic disease*", and "*informatics and new technologies in diabetic care*".

The **AMD Annals** represent a periodic publication that, since 2006, has allowed for the annual assessment of healthcare profiles for individuals with type 1 diabetes (DMT1) and type 2 diabetes (DMT2) under the care of Italian diabetology services. More in details, a widespread network of diabetology services, equipped with an electronic health record used for routine patient management, has a software provided by AMD that enables the extraction of a set of clinical information: the **AMD Data File**. The obtained dataset is used, by the association, for calculating quality-of-care indicators both at a centralized and local level.
In addition, the Annals' dataset also serves as a valuable source of observational research data. It has facilitated in-depth exploration of key aspects such as the care

[1] The information presented in this introductory part can be found here.

of elderly patients, gender medicine, cardiovascular, renal, and hepatic aspects, as well as the appropriateness of drug utilization.

As already reported in Chapter 1, AMD shared with Sapienza a series of documented files, representing the AMD real-world data, collected from the various Italian structures and containing information about diabetic patients from 2006 to 2018. These `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 (ATC) code [1] 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 part of the *data preparation* work presented in the next sections.

Also in the *AMD Data* context [19], the OBDM approach was exploited: by first modeling the ontology representing the content of AMD data, the domain knowledge is expressed and data can be properly loaded and prepared into a corresponding database. In this case, the technology used for storing the AMD data was **PostgreSQL**.

4.1 Ontology

The AMD ontology[2] has been formalised using the OWL formalism [4, 15] and consists of all the major relevant concepts and relations together with all the characterizing properties. Figure 4.1 contains a snippet of such an ontology.

Also in this case, the figures express the axioms of the ontology by `GRAPHOL`, the formalism that allows one to view the OWL ontologies in a diagram [52], as already outlined in Chapter 2. These are the conventions adopted:

- The concepts and attributes in red are those for which we do not have 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 4.1 we provided some example instances of the metadata concept `ExaminationType`, which represents all the different types of examination that are collected in the

[2] The appendix B contains the whole OWL 2 ontology's axioms.

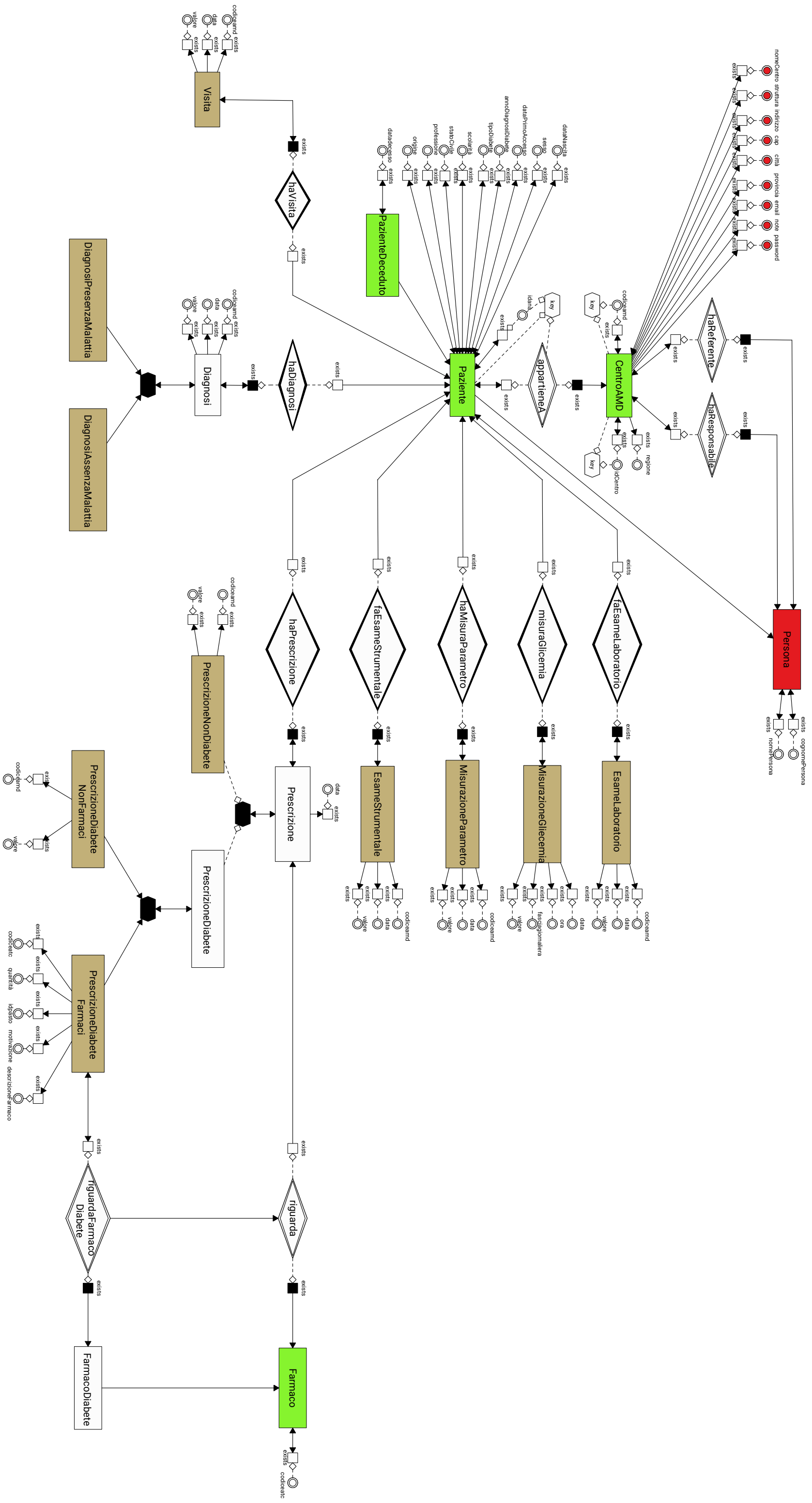

Figure 4.1. The ontology representing the main concepts and relations involved in the data published yearly by AMD, as well as an example of the metadata describing them.

dataset. In the example depicted, we show seven different types of examination, namely the asymptomatic hypoglycemia, the symptomatic hypoglycemia, the foot and GISED questionnaires, the specific and overall anamnesis, 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`, we represent 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`.

- 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 description of the main axioms of the modeled ontology is provided.

The data comes 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 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 `data 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`).

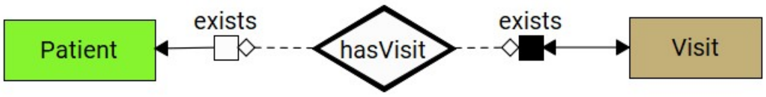

Figure 4.2. relation between `Patient` and `Visit`

- `Patients` are subject to diagnosis (`Diagnosis` concept and related sub-concepts).

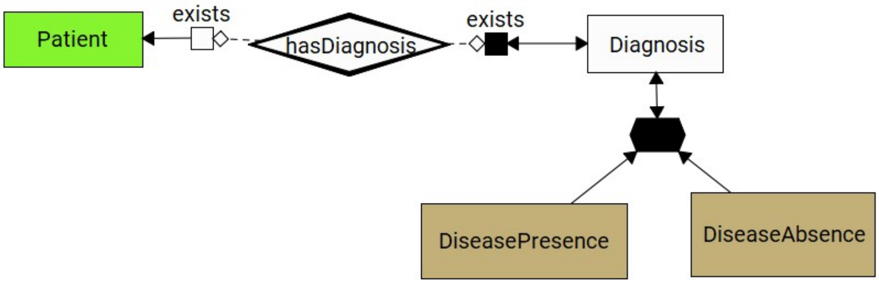

Figure 4.3. relation between `Patient` and `Diagnosis`

- Patients perform different types of exams (concepts Laboratory Exam, Blood Glucose Measurement, Parameter Measurement, Instrumental Exam).

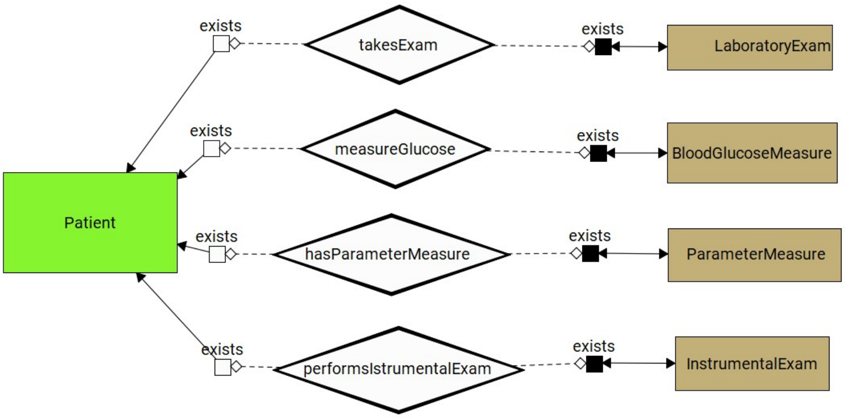

Figure 4.4. relations between Patient and other concepts

- Patients are assigned prescriptions by doctors (Prescription concept and its sub-concepts).

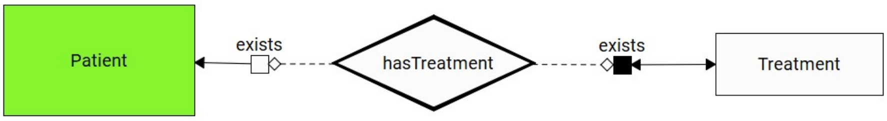

Figure 4.5. relations between Patient and Treatment

- Prescriptions can relate to drugs (Drug concept and related sub-concepts).

Despite the existence of DMTO ontology that models the *diabetes mellitus domain* as a whole [29, 30], 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.

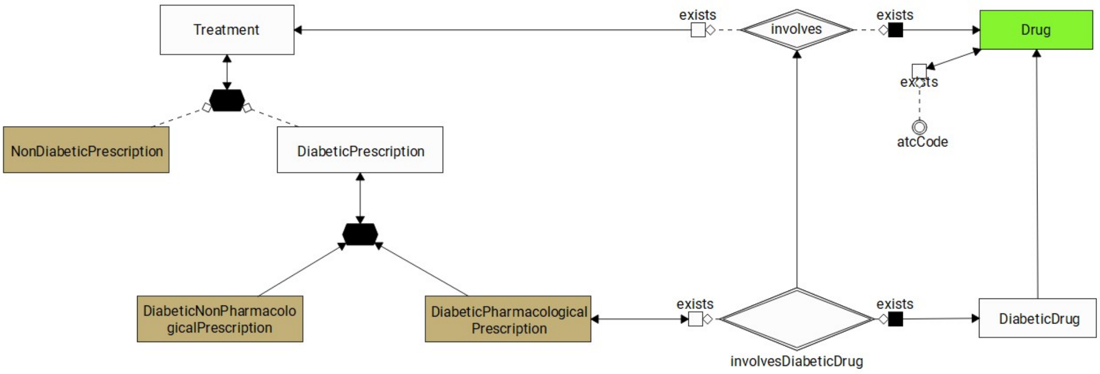

Figure 4.6. relations between `Treatment` and `Drug` and related sub-concepts;

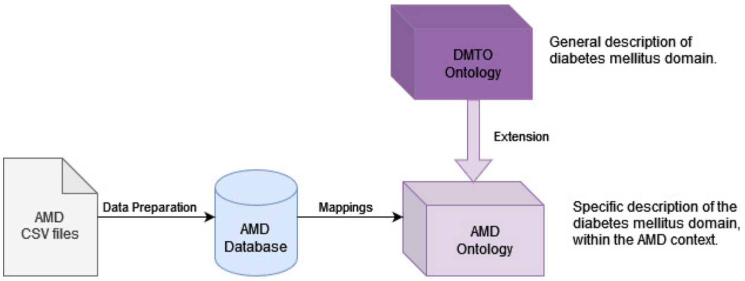
DMTO Ontology
General description of diabetes mellitus domain.
Extension
AMD CSV files
Data Preparation
AMD Database
Mappings
AMD Ontology
Specific description of the diabetes mellitus domain, within the AMD context.

Figure 4.7. 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.

4.2 Database

Following the same methodology presented in Chapter 1 of this dissertation, 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 metadata. The former part of the schema (data tables) was meant to contain the actual data, whereas the latter (metadata tables) contains metadata 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.

Data tables

The data tables are the following:

- `data.amdcenters`: registry of the centers; A list of the most important attributes follows:
  - centerid: *number identifying AMD centers in the current database*;
  - amdcode: *AMD code of the center (identifier for AMD)*;
  - region: *name of the region to which the center belongs.*
- `data.registry`: patient registry; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient*;
  - sex: *the sex of the patient*;
  - yeardiagnosisdiabetes: *the year in which diabetes was diagnosed for the patient*;
  - diabetestype: *the type of diabetes (values from 2 to 20 excluding 19) for the patient*;
  - educationlevel: *the level of education (unique for each patient)*;
  - maritalstatus: *the marital status (unique for each patient)*;
  - profession: *the profession of the patient*;
  - ethnicorigin: *the ethnic origin*;
  - birthdate: *the birth date of the patient*;
  - firstaccessyear: *the year of the patient's first access to the center*;
  - deceasedate: *the date of the eventual decease of the patient.*
- `data.visit`: records "visit" events; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient conducting the visit*;
  - date: *the date of the visit*;

  - amdcode: *the AMD code indicating the type of visit; the types of visits are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the amdcode.*

- `data.laboratoryparameterexam`: records laboratory examination or parameter detection events; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient conducting the examination or parameter measurement*;
  - date: *the date of the examination*;
  - amdcode: *the AMD code indicating the type of examination or parameter measurement; the types of examinations and measurements for this category are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the amdcode.*

- `data.glycemia`: records capillary glucose measurements performed by the patient at home; A list of the most important attributes follows:
  - centerid: *the identifier of the center in the extraction*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient conducting the visit*;
  - date: *the date of the blood glucose test*;
  - time: *the time of the blood glucose test*;
  - dailyrangecode: *code indicating the daily range of the test*;
  - smbglvl: *the value of the self-monitored blood glucose.*

- `data.instrumentalexam`: records instrumental examination events; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient who underwent the examination*;
  - date: *the date of the examination*;
  - amdcode: *the AMD code indicating the type of examination; the types of examinations for this category are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the type of examination, i.e., based on the value of amdcode.*

- `data.diagnosis`: records “diagnosis” events, which in addition to diagnoses related to diabetes, also include comorbidities and complications; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient subject to the diagnosis*;

  - date: *the date of the diagnosis*;
  - amdcode: *the AMD code of the diagnosis; the types of diagnoses are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the value of amdcode.*

- `data.diabetesdrugprescription`: records events of prescription of drugs as part of the treatment of diabetes; A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient subject to the diagnosis*;
  - date: *the date of the prescription*;
  - atccode: *the ATC code of the prescribed drug*;
  - drugdescription: *the name of the drug and dosage contained in the package*;
  - quantity: *the number of tablets or units of insulin prescribed to the patient*;
  - mealid: *code indicating the meal for drug intake.*
- `data.nondrugdiabetesprescription`: records prescriptions related to blood glucose control and non-drug related prescriptions (for example diet); A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient subject to the diagnosis*;
  - date: *the date of the prescription*;
  - amdcode: *the AMD code of the prescribed treatment; the types of treatments for this category are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the value of amdcode.*
- `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); A list of the most important attributes follows:
  - centerid: *the identifier of the center*;
  - patientid: *the identifier of the patient within the AMD center – together with centerid, it forms the identifier of the patient subject to the diagnosis*;
  - date: *the date of the prescription*;
  - amdcode: *the AMD code of the prescribed treatment; the types of treatments for this category are listed in the corresponding metadata table*;
  - value: *the meaning varies depending on the value of amdcode.*

Meta-data tables

The metadata 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 metadata 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), the classification system that organizes diseases and injuries into groups on the basis of defined criteria [60]. It has to be pointed 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, this project refers to the 9th release because it is the classification system used in the AMD dataset at the moment. However, since the presented approach is insensible to the classification version used, it will eventually be updated to more recent versions as soon as the AMD dataset will also be updated.

4.3 Data Preparation

In this section, the cleaning process carried out on the AMD dataset [19, 59] is presented. It is important to note that, in this project as it was for RIAP scenario presented in Chapter 3, 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 4.3.2 has 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 250 different centres for diabetes treatment in Italy since 2005[3]. 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 in the gathered data.

Many cleaning actions were necessary to overcome these problems. It is possible to distinguish the actions taken in three 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 it has been decided to adopt the second criterion since at this stage the unique goal was to keep each table consistent with itself and with its metadata descriptions.

4.3.1 Actions to resolve errors in the data

An important objective of the cleaning was to make data and metadata descriptions consistent with each other. Whenever a discrepancy between what was described in the metadata, and what was stored in the data occurred, two different strategies could be adopted: changing the data to make them consistent with their metadata description, or changing the metadata to effectively describe what was stored in the data. It has not been possible (neither was it 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 experts of the domain, i.e. the physicians of our group.

For example, the metadata 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 metadata 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 metadata for such laboratory exams have been corrected.

[3] It has to be clarified that not all the 320 centres constituting the AMD network included their data in the latest AMD annals.

4.3.2 Actions to resolve missing and incomplete data

As one might expect from a dataset of this magnitude and complexity, considering the challenges mentioned in the literature and also the case study presented in Chapter 3, 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 for these types of tuples were found:

- positive disease diagnosis with associated a `NULL` value;
- positive disease diagnosis with associated a value representing the ICD-9-CM [60] 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 [60] of a disease diagnosis is obviously an added value. Therefore, we removed all `NULL` values 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.

4.3.3 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, being another challenge detailed by the literature on data management in healthcare. This is the reason why the following actions to mitigate the risks of identity disclosure were taken:

- to remove the specific day and month from the date of birth of patients, by only leaving the more general year of birth.
- to remove the specific day and month from the date of first access to a center in the AMD network.
- to remove the specific day and month from the date of first diagnosis of diabetes.

- to substitute 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 were collected into one single bigger fictitious center. As a result, it was checked that all patients that were recognizable in their original center were indistinguishable from other patients in the new center.

4.4 Data analytics on AMD data

Once the database reached a consistent state, several research groups, by leveraging machine learning approaches and the potential represented by AMD data, started to focus on different tasks. In this section, after a brief discussion on the effectiveness of data preparation discussed so far, one of the mentioned research targets will be explored: *patient clustering based on her temporal trend of Glycated Hemoglobin (HbA1c)*. In fact, 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, it is crucial to investigate the trends of HbA1c over the years.

4.4.1 Effectiveness of data preparation

To illustrate the improvements resulting from the data preparation process discussed earlier, consider the relatively simple task of retrieving all patient demographics along with their HbA1c value sequences across the years, which involves only a subset of the AMD dataset.

To compare the performance of this task between a database directly derived from data sources and the stable version of the AMD database obtained through ontology conceptualization and data cleaning, another relational database was created, in this case perfectly mirroring the content of the raw data. In the scenario of the source-based database, personal data is split into three tables, while HbA1c sequences are extracted through an aggregation query from three tables of numeric information. In the latter setting, the AMD database, only one table holds patient demographics, and another contains the required HbA1c measures.

Beyond analyzing query complexity qualitatively, the subsequent step involves executing both queries (with identical semantics) and investigating potential differences between the two settings.

This task example revealed significant disparities between the *source-based database* and the *AMD database* resulting from the discussed data preparation.

Comparing the number of records, it was found that **16.8%** are missing in the final database due to data cleaning, which eliminated duplicates or incorrect records. This indicates that a substantial portion of records in the source-based database contained inconsistencies.

However, the most notable result involves the value of the clinical parameter. Analyzing the mean of HbA1c sequences for each patient and computing the average of these mean values yielded a plausible outcome (7.48%) in the final database, aligning with clinical literature guidelines. In contrast, the `csv` files produced an inconsistent outcome (121,394,866%) due to records containing both plausible and out-of-scale, and thus inconsistent, HbA1c values.

These findings suggest that querying `csv` data sources without data preparation can yield results with numerous inconsistencies, requiring collaboration with domain experts for resolution. The data preparation process discussed here is general and context-oriented, ensuring consistency across the entire database, unlike an ad hoc and task-dependent data preprocessing.

4.4.2 Clustering of diabetic patients

The study of HbA1c is looking for a trajectory-based clusterization of AMD patients. To reach this objective, from the total group of DMT2 patients, a discussion between computer science engineers and physicians brough to the definition of the following preconditions for each patient to be selected:

- *the patient has T2DM;*
- *the patient should have HbA1c measures for a span of **10** years*;
- *the number of the measure should be at least **10***;
- *the distance in time between one measure and the subsequent one could not be greater than **1** year.*

From the 1,174,908 records (i.e., diabetic patients with HbA1c measures stored in the database), *68,486* passed the filtering defined above. After a more in-depth discussion with the stakeholders, it has been also decided a further criteria of filtering:

- **neodiagnosis condition**: *the distance between the diagnosis of diabetes and the first HbA1c measure (valid for the study) should be at most **5** years.*

This last condition selected only *20,376* patients from the *68,486.* However, the subsequent analyses have been carried on both the cohorts (with and without considering the neodiagnosis condition). From this moment on, this dissertation will refere to the *68,486* patients as *complete cohort (CC)* and to the subset of CC which fulfilled the neodiagnosis condition as *neodiagnosis cohort (NC).*

Once identified the two portions of data, which were compliant with the stakeholders' requirements, the subsequent step was to find a way of finding possible sub-groups with relevant differences.

The strong belief of the physicians involved in this project was that a *sort of natural clusterization* of the diabetic patients should exist. This statement was primarily corroborated by the clinical practice, and it has been also addressed by some previous research [6, 50].

Therefore, the decision was to try to investigate the presence of those subgroups exploiting an *unsupervised machine learning approach*: **K-Means algorithm**. The description of the classical K-Means algorithm is provided as a pseudocode snippet:

```
Input: $k$ integer, $D$ datapoints
Choose randomly $k$ centroides: $c_1, ..., c_k \in D$
$i := 0$;
while $i <= N$ do
    for $x_i \in D$ do
        for $c_j \in (c_1, ..., c_k)$ do
            $d_{(i,j)} := distance(x_i, c_j)$;
        end for
        $cl(x_i) \leftarrow c_j \mid j := min_j(d_i)$;
    end for
    for $c_i \in (c_1, ..., c_k)$ do
        $c_i = average_{\{x|cl(x)=c_i\}}(x)$;
    end for
    $i \leftarrow i + 1$
end while
```

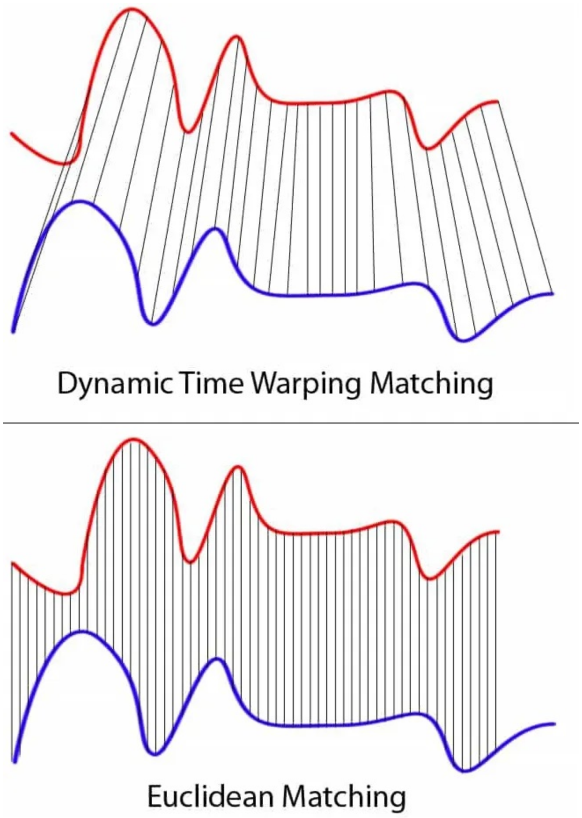

Figure 4.8. The difference between the eucledian distance and the dynamic time warping distance in capturing curves' changes over time (available here.)

The algorithm chooses k centroids *randomly*. Then, until convergence, assigns to each record in the dataset one of the centroids. This is done selecting the **minimum distance** between the record and each centroid. Subsequently, it re-computes the centroids, averaging the intra-cluster distances between the records assigned to each cluster.

Within this algorithm, the most important role is played by the `distance` function, allowing both to assign records to a cluster and to re-compute the distance among the assigned elements. The classical formulation of K-means algorithm uses the *Euclidean Distance* between the features on which the clustering has to be performed. However, in the case of the HbA1c measures, the case was slightly different. The case was not the one of a set of different features, on which it was required to compute the distance, but the same feature varying along a time period of ten years (i.e., multiple values of the same feature).

This is the reason why the selected metric for the `distance` function was the so-called: **dynamic time warping (DTW)**:

$$DTW(x_1, x_2) = \min_\pi(\sqrt{\textstyle\sum_{(i,j)\in\pi}(x_{1i} - x_{2j})})$$

The distance between two records, in this case, is given by the minimum square root of all the paths π of (i,j) couples. Therefore, in our case, π represents all the possible combination of subsequent measures of HbA1c measures for patient x_1 and patient x_2. The minimization is thus *on the shape of the two curves*. For example, if patient

x_1 has a decrease of HbA1c after 2 years and x_2 has a very similar decreasing, but after 4 years, this distance metric can *detect* this similarity. On the other hand, the *Euclidean Distance*, being point-wise, would not have taken into account that fact.

Once the algorithm was chosen, it was carried out an analysis, through the *Elbow Method*, in order to have a clue on the *number of clusters*. The result of the application of this methodology (shown in Figure 4.9) suggested that k = 4 is the best choice as input for the Temporal K-Means.

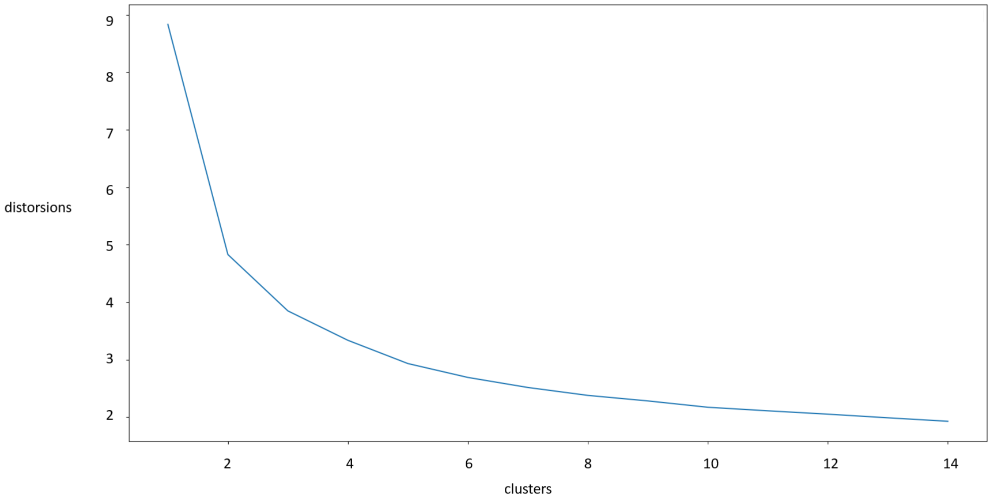

Figure 4.9. Results of the elbow method, applied to the NC.

Therefore, the two cohorts (CC and NC) and k = 4 were passed as parameters to the TKM. After returning the four clusters, their main characteristics (in term of average HbA1c per year) were shown to the domain experts. Here is the list of the four clusters (also shown graphically in Figure 4.10), with a qualitative description provided by the physicians:

- **compensated**: remaining, along the 10 years, always within controlled values of HbA1c;
- **moderately decompensated**: patients fluctuating, along the 10 years, from compensated values of HbA1c to moderately uncompesated ones;
- **decompensated with continuous improvement**: patients switching from an initial phase of severe uncompensated values of HbA1c to a more controlled situation in the subsequent years;
- **decompensated**: remaining, along the 10 years, always within decompensated values of HbA1c.

In what follows, the behaviour of the obtained clusterization is discussed, analyzing some of the most important patient-related characteristics. In particular, the subsequent analyses had the target of answering the following questions:

- **sex**: *is the clustering suggesting a different response to the diabetes treatments between two genders?*

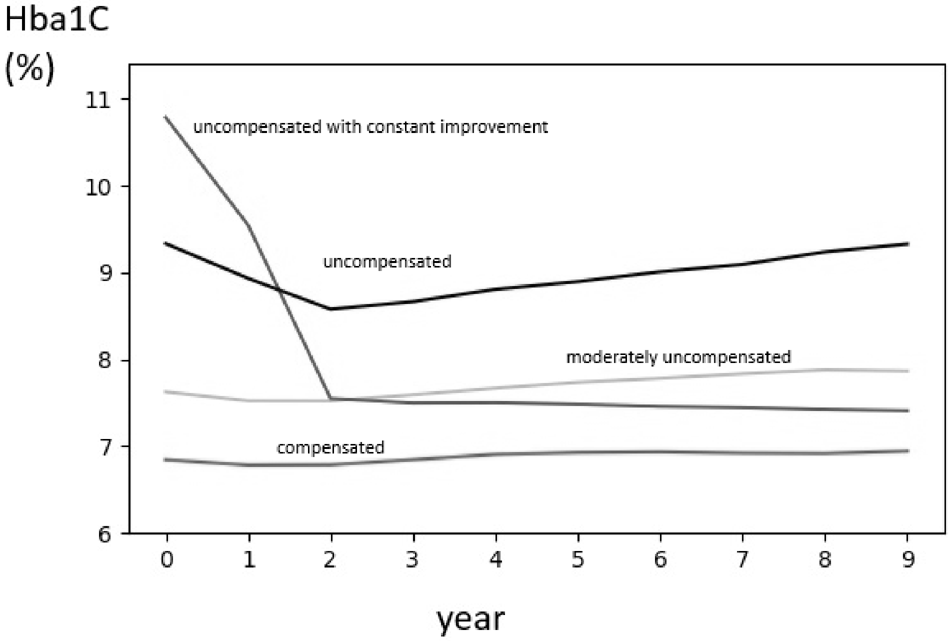

Figure 4.10. The average value of HbA1c for the 10 years of the four NC clusters

- **age**: *if taking into account the age a patient was registered into the system, are there any differences with respect to the average age within the different clusters?*
- **BMI**: *is the body mass index influencing the response of a patient to the diabetes treatments?*
- **medications**: *do the different clusters reflects (or are reflecion of) different types of diabetes treatments (insuline vs other drugs) prescribed to the patients?*
- **complications**: *do two diabetes-related complications, namely nephropathy and retinopathy, have some kind of influence on (or are influenced by) the cluster assigned to a patient?*

4.4.3 Demographics and BMI

The demographics which are taken into account are sex and age, while the body mass index is taken into account as a relevant clinical measure of the patients' habits. The figures shows the percentage difference between the distribution in the sample (NC or CC) and the considered cluster for sex, age at first measure, bmi, respectively.

Sex

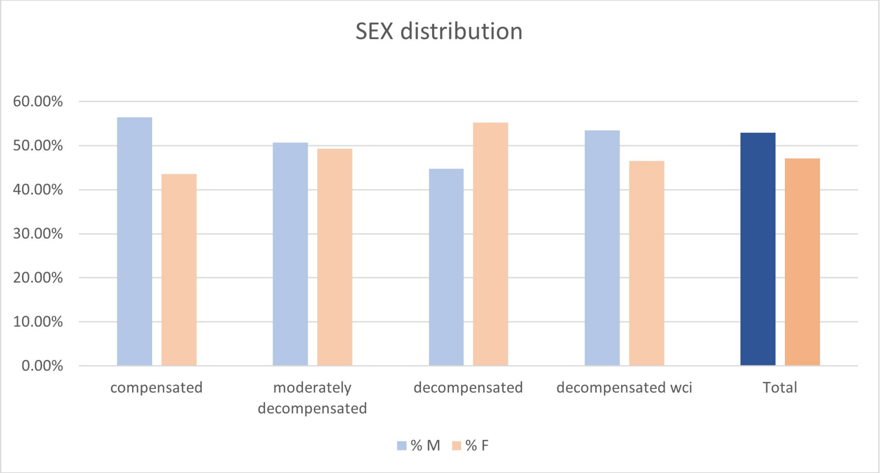

Figure 4.11. Sex distribution per cluster (CC)

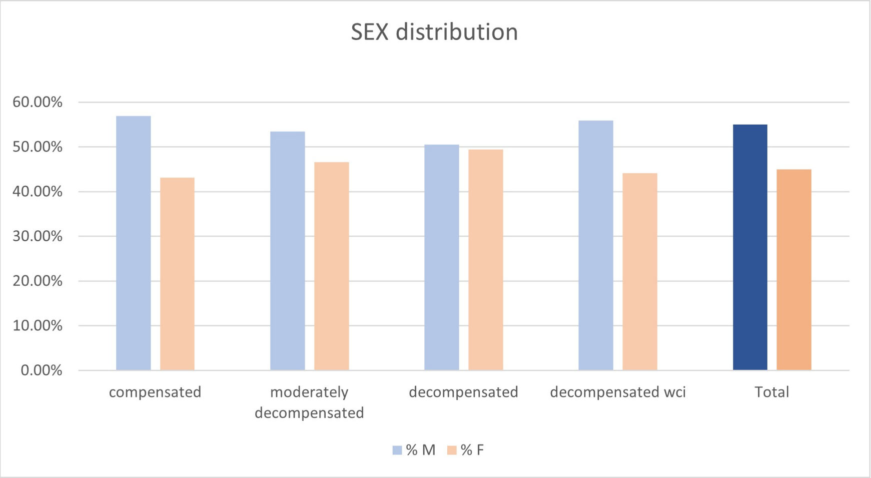

Figure 4.12. Sex distribution per cluster (NC)

If looking at the clusters behaviour with the respect to the *sex distribution*, it is possible to initially assert that the gap between males and females is **10%** for NC, being **5.91%** for the CC. In both cohorts, this difference from the gap is higher in *compensated* cluster (CC: +6.91%; NC: +3.75%), while lower in the *decompensated* (CC: -16.41%; NC: -8.90%). The *moderately decompensated* cluster is between the two extremes (CC: -4.59%; NC: -3.16%). The *decompensated wci* roughly resables the *overall distribution* (CC: -1.10%; NC: -1.77%).

Age

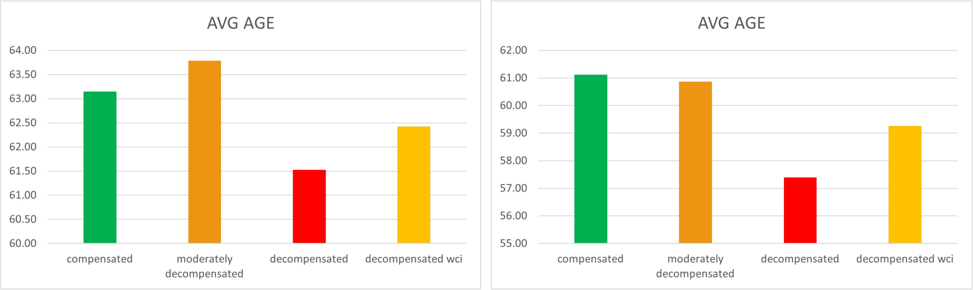

Figure 4.13. Average age per cluster (CC, NC)

The average **age**, shown in Figures 4.13, for the *compensated* (CC: 63.15; NC: 61.12) and the *moderately decompensated* (CC: 63.78; NC: 60.87) clusters is higher with the respect to the *decompensated wci* (CC: 62.42; NC: 59.26) and the *decompensated* (CC: 61.53; NC: 57.39) ones.

As expected, if comparing the two cohorts, it is possible to observe that the age at first considered measure for NC is *lower on average*.

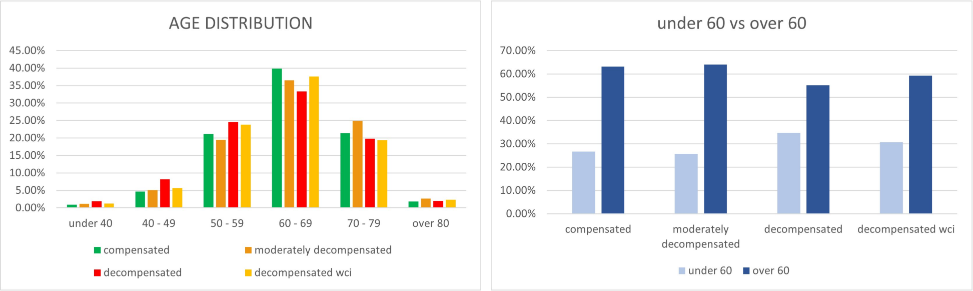

Figure 4.14. Age distribution per cluster (CC)

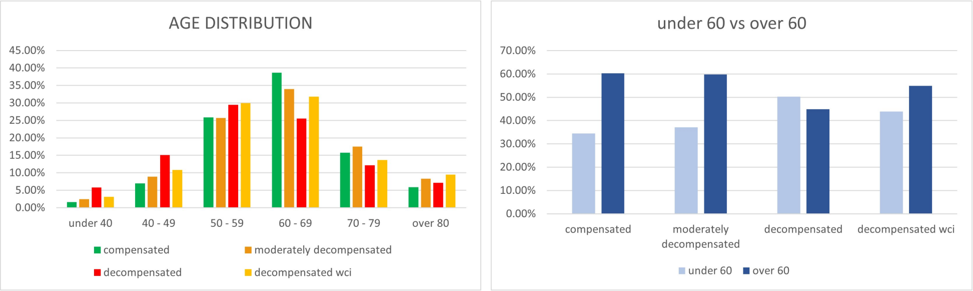

Figure 4.15. Age distribution per cluster (NC)

If considering the distribution of age inter-clusters, shown in Figures 4.14 and 4.15, the presence of **under 60** patients within the *compensated* (CC: 26.66%; NC: 34.42%) and *moderately uncompensated* (CC: 25.75%; NC: 37.06%) is less prevalent than in the *decompensated wci* (CC: 30.72%; NC: 43.88%) and *decompensated* (CC: 34.68%; NC: 50.33%).

Body Mass Index (BMI)

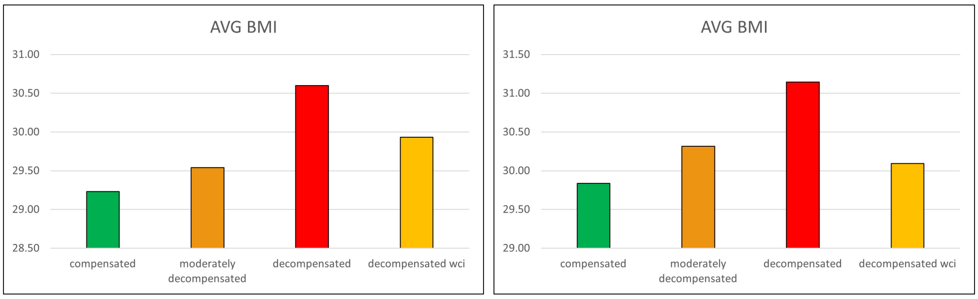

Figure 4.16. Average age per cluster (CC, NC)

The average **bmi**, shown in Figures 4.16, is lower in the *compensated* (CC: 29.23; NC: 29.83) cluster, higher in the the *decompensated* (CC: 30.60; NC: 31.14) cluster, and it is the middle between the two extreems within the *moderately decompensated* (CC: 29.54; NC: 30.32) clusters is higher with the respect to the *decompensated wci* (CC: 29.93; NC: 30.09) and ones.

If comparing the two cohorts, it is possible to observe that the BMI at for NC is *higher on average in all the clusters* with respect to the CC.

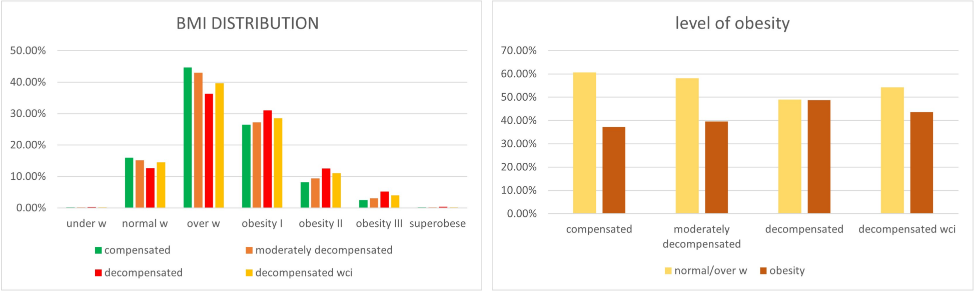

Figure 4.17. BMI distribution per cluster (CC)

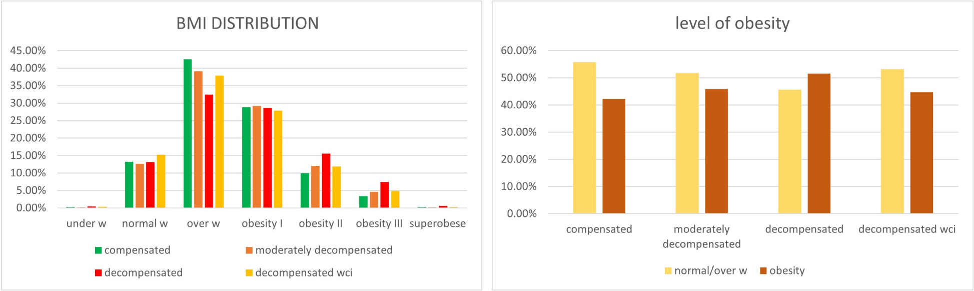

Figure 4.18. BMI distribution per cluster (NC)

If considering the distribution of BMI inter-clusters, shown in Figures 4.17 and 4.18, the presence of **obesity** patients within the *compensated* (CC: 37.25%; NC: 42.16%) cluster, more prevalent within the **decompensated** patients (CC: 48.75%; NC: 51.60%), and it is in a middle-range within *decompensated wci* (CC: 43.58%; NC: 44.66%) and *moderately decompensated* (CC: 39.63%; NC: 45.84%) clusters.

4.4.4 Medications

A further and deeper investigation is represented by the analysis of the pharmaceutic treatments of the two patients' cohorts (namely, CC and NC). In particular, the aim of this part of the study is to assess whether there were substantial differences inter-clusters if taking into accounts the drugs prescribed by the clinicians.

This kind of information is stored in the database through **ATC codes**[4]. As for ICD9-CM taxonomy, the codes are substantially represented in a *tree* form: the more one moves from a coarse to a fine grane, the longer is the length code.

The two main codes representing the most relevant cathegories of pharmacologic therapies are the prescription of either **hypoglycemic drugs** (A10B) or **insuline** (A10A) and the corresponding sub-branches.

The most interesting results are shown by Figures 4.19 and 4.20. They show that the more the cluster reflecs the decompensation, the more probable is the assumption of insuline and, on a temporal basis, the `decompensated wci` cluster, if compared with the `moderately decompensated`, has a worse baseline behaviour (higher assumption of insuline within the cluster), but a better behaviour at the end of the observation, when the percentage of patients with an insuline therapy is lower.

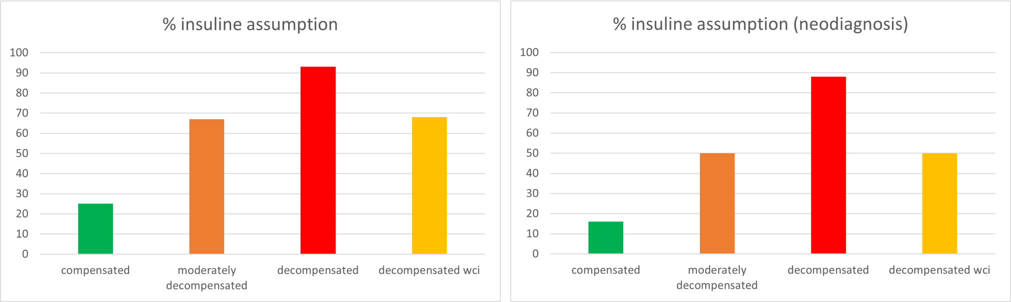

Figure 4.19. Insuline assumtpion % per cluster.

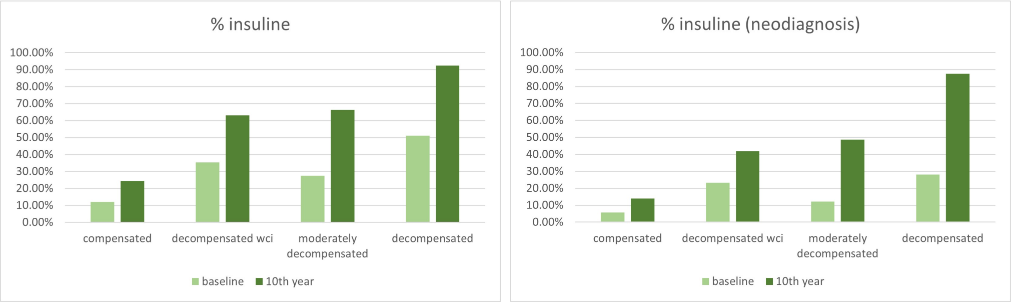

Figure 4.20. Insuline assumtpion % per cluster, baseline versus 10th year.

This analysis on the therapies seems to perfectly reflect the qualitative nature of the clusters, thus validating them from this perspective. The two extremes clusters (`compensated` and `decompensated` are at two opposite sides, while the two intermediate clusters (`moderately decompensated` and `decompensated wci`) have a similar average behaviour, with different evolutions of their trends.

[4] citare ATC

4.4.5 Complications: nephropathy and retinopathy

The final study performed on the four clusters was the prevalence of two common T2DM comorbidities, namely **nephropathy** and **retinopathy**. Due to the nature of the dataset and the complexity in defining the two diseases, this part of the study required a further reduction of the number of patients.

Nephropathy

Before showing the results about the prevalence of nephropathy within each cluster, the algorithm for the definition of the diagnosis of the disease is described in details.

1. **Start point query** to retrieve `idcenter`, `idana`, `gender`, and `date` of the first glycated measurement for patients who:
   - Belong to the CC cohort;
   - Have measurements of GFR (STITCH05) and ACR (AMD023, AMD024, AMD026, AMD111, AMD910, AMD911) on the same date.
2. The following thresholds have been considered:
   - **GFR** (Code STITCH05), the threshold is *59*;
   - **ACR**
     - Code AMD023: the threshold is *3*;
     - Code AMD024: the threshold is *20*;
     - Codes AMD111, AMD026, AMD910: the threshold is *30*;
     - Code AMD911: the threshold is *2.5* for men, *3.5* for women.

   In case of discordant measurements (about 33,000) of ACR on the same date (GFR measurements on the same dates are *never* discordant), only the GFR with its threshold was considered to determine nephropathy = 1 or nephropathy = 0.
3. From this dataset, patients were considered pathological (1) if, from a certain year onwards, they *always* had the value 1 on the dates of the measurements taken, and non-pathological (0) if they *never* had a value of 1 on the dates of the measurements taken, *or* if they had any moment when they became pathological, then returned to normal values (temporary nephropathy, considered as *absence* of nephropathy).
   dataset: `idcenter`, `idana`, `nephropathy` (0 or 1), `year` (year of the start of observations on nephropathy), `year_onset` (year of onset of non-temporary nephropathy or 0), `year_onset_temporary` (year of onset of temporary nephropathy or 0);
4. From the dataset, those patients with less than 2 events of nephropathy diagnosis within the 10 years of observation, or with the last event occurred before the year 5, were excluded.

The number of patients of with presence or absence of nephropathy during the 10 years of follow-up was 59,842/68,486 (87.38%) within CC and 17,875/20,376 (87.73%) within NC.

The results of the nephropathy prevalence, pictured in Figure 4.21, show that the the difference between `decompensated` and `compensated` patient is more than

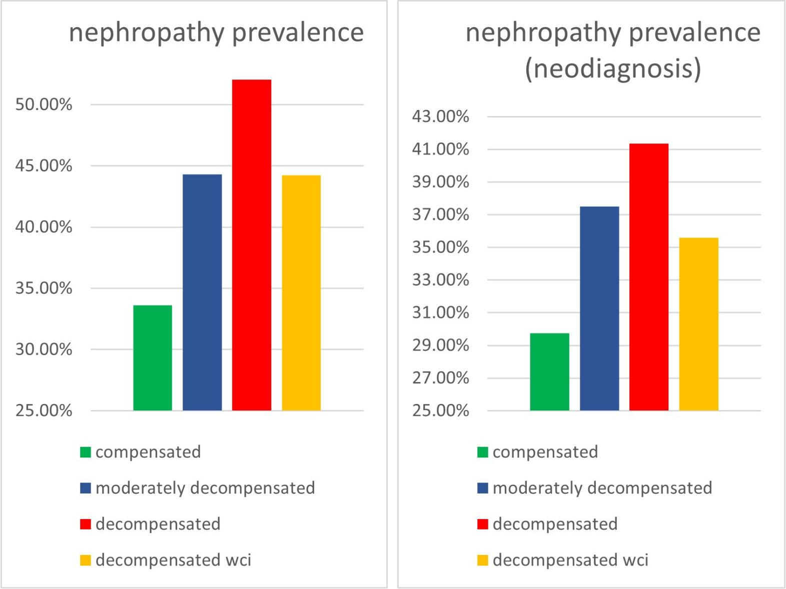

Figure 4.21. Nephropathy prevalence % per cluster.

15% for the entire dataset and above *10%* for neodiagnosis. The other two clusters are in the middle.

Retinopathy

Also for retinopathy, before showing the results about its prevalence within each cluster, the procedure for the definition of the diagnosis of the disease is pictured.

1. **Start point query** to retrieve retrieve `idcenter`, `idana`, `date` of the first glycated measurement for patients who:
   - Belong to the CC;
   - Have any diagnosis of retinopathy (AMD053, AMD054, AMD055, AMD056, AMD057, AMD204, AMD205, AMD210, AMD300);
   - Have a diagnosis of the absence of retinopathy (AMD130);
   - Have undergone screening for complications (AMD051, AMD052, AMD135).
2. For those patients who, in the first query, presented 2 (or more) events listed (diagnoses, "non" diagnoses, or screenings) on the same date, the following decisions were made:
   - At least one fundus examination (P) → the patient has retinopathy (retinopathy = 1)
   - At least one diagnosis of retinopathy (at any stage) → the patient has retinopathy (retinopathy = 1)
   - All other cases → the patient does not have retinopathy (retinopathy = 0)

   Therefore, for example, patients with a group of exams < N, P, N, .. > have retinopathy; patients with exams and diagnosis < N, N, AMD0XX > have retinopathy; patients with < P, AMD130, .. > have retinopathy, etc. Without these simplifications, thousands of observations (about 27,000) would be automatically excluded from the group.

3. At this point, patients who had a 0 (absence of retinopathy) on a date later than the one with a 1 were excluded (4705 patients, 7.18%).
   dataset: `idcenter`, `idana`, `retinopathy` (0 or 1), `year` (onset of retinopathy or start of observation for those with 0), `year_end` (end of observation, last diagnosis, or screening);

4. From the dataset, those patients with less than 2 events of retinopathy diagnosis within the 10 years of observation, or with the last event occurred before the year 5, were excluded.

The number of patients of with presence or absence of retinopathy during the 10 years of follow-up was 53,965 /68,486 (78.80%) within CC and 16,654/20,376 (81.73%) within NC.

The results of the retinopathy prevalence, pictured in Figure 4.22, show that `decompensated` cluster has the higher value (64.18% for CC and 40.44% for NC), followed by `decompensated wci` (49.49% for CC and 31.43% for NC), `moderately decompensated` (39.95% for CC and 22.58% for NC) and, finally, `compensated` (23.34% for CC and 15.37% for NC).

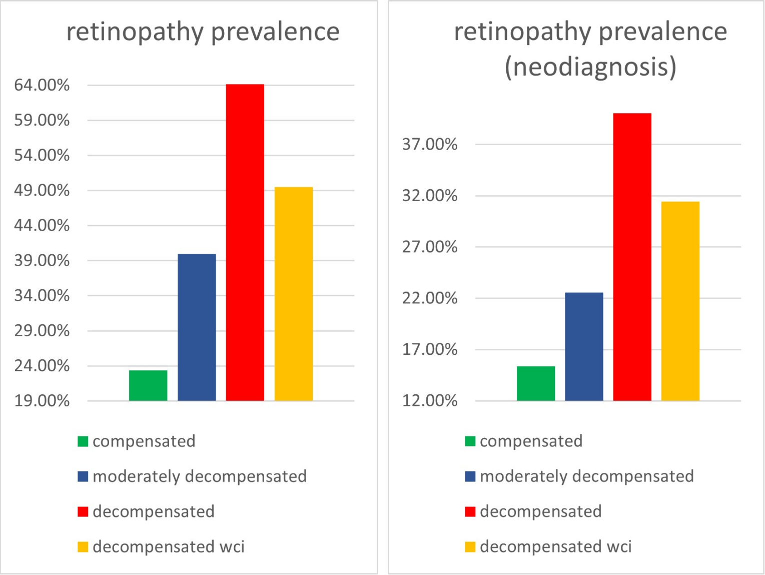

Figure 4.22. Retinopathy prevalence % per cluster.

4.5 Discussion

In this chapter, the OBDM approach was exploited to enhance the AMD-STITCH project, thus organizing, cleaning, and transforming the AMD dataset into a substantial shared asset, composed by real-world data, for all forthcoming data analytics tasks in the scenario of diabetes research.

Although considerable attention has been directed, by AMD, towards improving EMR software to facilitate manual data entry by physicians, multiple iterations of the cleaning process were necessary before achieving a stable state for the database. It is evident that future processing actions may be required, driven by the specific needs of upcoming data analytics tasks and inevitable corrective maintenance for datasets of this magnitude. Additionally, the possibility of AMD releasing updated versions of its dataset implies the need for further updates, including extensions to the data preparation outlined earlier.

It is possible to state that, also in this case, the application of the previously mentioned methodology represented a success. Despite the limitations given by the inconsistencies emerged from the data, the presence of the *AMD database* enabled to start five distinct tasks, leveraging the refined AMD dataset. While the data preparation work was promising, the true challenge lied in assessing its effectiveness in supporting data analytics. Therefore, this chapter also presented one of the research tasks carried on on those data, thus representing an example of **Computational Phenotyping Approach**, theoretically framed in Section 1.7.2. The results of this unsupervised machine learning technique reflect a possible answer to the clinical research target of better characterizing the T2DM phenotypes. This, beyond having a value *per se*, holds also a significant validation of the applied methodology.

In order to remark a substantial difference between AMD and RIAP projects, it is important to highlight the fact that, within AMD, the explored data analytics task, validating the efforts of applying the OBDM approach, is *data-driven.* Albeit needing further research, this suggests this methodology could be effectively employed to tackle the challenge arising from the literature analysis about AI in Healthcare, offering a solid method to enhance machine learning capabilities of being really successfully employed in clinical practice.