diabetesresearchandclinicalpractice 226 (2025) 112214

Contents lists available at ScienceDirect

Diabetes Research and Clinical Practice

journal homepage: www.journals.elsevier.com/diabetes-research-and-clinical-practice

# Characterizing the clinical profile and prevalence of people with diabetes attended in the hospital setting by using unstructured healthcare data and natural language processing: the Diabetic@ study

A.J. Blanco-Carrasco [a,*,1], J.F. Merino-Torres [b,*,1], M. Rubio Almanza [b], G. Canovas Molina [c], M.A. Brito-Sanfiel [d], D.E. Barajas Galindo [e], L.A. Cuellar Olmedo [f], D. Mauricio [g], S. Tofé Povedano [h], J.A. Balsa Barro [i], J.J. Aparicio Sánchez [j], M. Sequera Mutiozabal [j], B. Pimentel [j], A. Pérez Domínguez [j], C. Arias-Cabrales [k], V. Fanjul [k], L. Pérez de Isla [l,1], J.F. Navarro González [m,n,o,p,1]

[a] *Endocrinology and Nutrition Department, Hospital Clínic de Barcelona, Spain*
[b] *Endocrinology and Nutrition Department, Department of Medicine, Hospital Universitari i Politècnic La Fe, University of Valencia, La Fe Health Research Institute, Valencia, Spain*
[c] *Endocrinology and Nutrition Department, Hospital Universitario de Fuenlabrada, Madrid, Spain*
[d] *Endocrinology and Nutrition Department, Hospital Universitario Puerta de Hierro, Madrid, Spain*
[e] *Endocrinology and Nutrition Department, Hospital Universitario de León, León, Spain*
[f] *Endocrinology and Nutrition Department, Hospital Universitario Río Hortega, Valladolid, Spain*
[g] *Endocrinology and Nutrition Department, CIBERDEM, Hospital Universitari Santa Creu i Sant Pau, Barcelona, Spain*
[h] *Endocrinology and Nutrition Department, Hospital Universitari Son Espases, Mallorca, Spain*
[i] *Endocrinology and Nutrition Department, Hospital Universitario Infanta Sofía, Madrid, Spain*
[j] *Cardiovascular, Renal and Metabolism Medical Department, AstraZeneca, Spain*
[k] *SAVANA Research Group, S.A, Spain*
[l] *Cardiology Department, Hospital Clínico San Carlos, Madrid, Spain*
[m] *Research Unit and Nephrology Department, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain*
[n] *RICORS2040 (RD24-0004-0022), Instituto de Salud Carlos III, Madrid, Spain*
[o] *Instituto de Tecnologías Biomédicas. Universidad de La Laguna, Tenerife, Spain*
[p] *Facultad de Ciencias de la Salud, Universidad Fernando Pessoa Canarias, Las Palmas de Gran Canaria, Spain*

ARTICLE INFO

*Keywords:*
Diabetes Mellitus (DM)
Natural Language Processing (NLP)
Machine Learning (ML)
Unstructured Data
Diabetes Comorbidities
Diabetes Complications

ABSTRACT

*Aims:* This study aimed to evaluate the potential of unstructured electronic health records (EHRs) data, analyzed using natural language processing (NLP) and machine learning (ML), to describe the prevalence and clinical spectrum of diabetes mellitus (DM) in hospitals.

*Methods:* A multicenter, retrospective study was conducted using EHRs from eight Spanish hospitals (2013–2018). Unstructured data were extracted using EHRead® (NLP and ML) and SNOMED CT. Individuals with type 1 or 2 DM (T1DM/T2DM) were identified, and a semi-supervised ML classifier was developed for unregistered types (UrDM). DM prevalence and related complications were analyzed in the final subpopulations (sT1DM/sT2DM).

*Results:* From 56,181,954 EHRs of 2,582,778 individuals, 638,730 were identified with DM: 75.4% with UrDM, 21.3% with T2DM, and 3.3% with T1DM. The ML model reclassified 93.5% as T2DM and 6.5% as T1DM. Over 50% of relevant variables like anthropometrics, laboratory values and treatments were missing. The prevalence of sT1DM/sT2DM was 2.6%/38.4%. Major comorbidities included hypertension, dyslipidemia, chronic kidney disease (CKD), ischemic heart disease, and chronic heart failure (CHF). CKD and CHF were the most frequent complications for sT1DM/sT2DM at 60 months.

* Corresponding authors.
*E-mail addresses:* ablanco@clinic.cat (A.J. Blanco-Carrasco), merino_jfr@gva.es (J.F. Merino-Torres).
[1] These authors contributed equally.

https://doi.org/10.1016/j.diabres.2025.112214
Received 20 August 2024; Received in revised form 31 March 2025; Accepted 29 April 2025
Available online 2 May 2025
0168-8227/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Conclusions:* NLP and ML for profiling DM using EHRs unstructured data are helpful, but additional data and better EHR documentation are crucial.

## 1. Introduction

Diabetes mellitus (DM) is a severe, chronic metabolic disease that represents a major public health problem [1]. Contributing factors such as an aging population, sedentary lifestyles, and unhealthy diets have led to a global increase in DM cases and prevalence, now considered an epidemic [2]. The International Diabetes Federation Diabetes Atlas 2021 reports a global DM prevalence of 10.5% in adults aged 20–79 years, with the highest rate of 24.0% in those 75–79 years old [3]. Notably, the proportion of people with undiagnosed disease worldwide was estimated to be 44.7%. In Spain, data from the nation-wide cohort Di@bet.es study reported a prevalence of DM of 13.8% (95% CI, 12.8 to 14.7%) [3–4]. Due to hyperglycemia and individual components of the metabolic syndrome, subjects with diabetes have a high burden of comorbidity and an increased risk of cardiovascular events, leading to reduced life expectancy [5–6].

Although the rates of certain major DM-related complications —such as myocardial infarction, stroke, and amputation— are decreasing due to improvements in disease management, it continues to impose a substantial burden in terms of associated issues, which have a significant impact on hospitalization rates, length of stay and healthcare costs, and are a major cause of disease-related morbidity and mortality [5].

This underscores the need for improved control measures and a deeper understanding of disease progression, particularly in the population attended in hospital settings who face a higher risk of complications but are rarely studied [7]. Moreover, it is important to have up-to-date epidemiological data which are crucial for effective patient care and prevention strategies [8,9]. However, there are several challenges for obtaining this information such as the lack of comprehensive registries on disease prevalence and outcomes, and because most of the clinical information is currently collected as unstructured data. This scenario hinders the exploitation of DM data using standard study designs with structured data or claims and usual statistical methods.

In addition, although clinical guidelines categorize diabetes mellitus into type 1 (DM1; caused by β-cell destruction), type 2 (DM2; driven by insulin resistance), and other less common types (e.g., hereditary variations, diseases of the exocrine pancreas, gestational diabetes, pharmacological or chemically induced diabetes) [9], evidence suggests significant misclassification and incomplete coding in routinely collected data [10]. This can result in cases where the specific type of diabetes is not clearly defined or easily identifiable in patients' electronic health records (EHRs) [11], leading to the classification of these patients as having an unrecorded diabetes type (UrDM). Such changes can substantially impact the analyses and conclusions drawn from real-world data (RWD) studies [12], emphasizing the need for advanced tools to accurately identify and address these cases within the data [13][14].

In this context, natural language processing (NLP) and machine learning (ML) techniques hold great potential to bridge this gap, thanks to their potential for the extraction and analysis of large amounts of unstructured, free text from electronic health records (EHRs) in a promptly manner [15,16]. Unlocking the narrative content within EHR yields nuanced clinical insights that transcend the limitations of structured data. Despite the rapidly growing number of applications of NLP tools focused on DM research, very few studies —less than 20% in a recently published systematic review— have reported results with potential implications for improving DM care [17].

In this real-world data (RWD) study, we aim to use NLP and ML techniques to analyze only unstructured healthcare data coming from free-text narratives from a broad population of individuals attended in a hospital setting to characterize the prevalence and clinical spectrum of DM. This analysis will help us better understand this high-risk and complex population, shedding light on the potential for leveraging such data.

## 2. Research design and methods

### 2.1. Study design and data source

This is a multicenter, retrospective, observational study based on RWD by analyzing clinical information extracted from the unstructured free text of individuals' EHRs attended in the participating hospitals. Data were collected from all available departments in each participating site (including inpatient hospital, outpatient hospital, and emergency service) between 1st January 2013 and 31st December 2018. The study was conducted in 8 hospitals, located in 5 regions, from the Spanish National Healthcare Network (Supplementary Table 1).

### 2.2. Study population

The source population included all individuals attended in the participating hospitals during the study period, including both inpatients (hospitalized) and outpatients (ambulatory). Subjects with a confirmed DM diagnosis (at least one unambiguous, non-negated mention of 'DM' in their EHRs, considering all alternative names) were selected. Individuals were excluded if they had DM only registered in the family history or a registered type of DM other than type 1 or type 2, such as gestational diabetes. Subjects with inconsistent 'DM' affirmation were also excluded. This refers to individuals whose EHRs entries either included speculative mentions of DM (e.g. "possible DM" or "suspected DM") or contained contradictions within the same patient's records such as an initial affirmative mention followed by a subsequent negation, like "DM ruled out".

After selection, individuals were assigned a DM type as reported in the unstructured data as T1DM, T2DM or unregistered type of DM (UrDM). Only persons with T1DM aged $\geq$ 2 years and persons with T2DM aged $\geq$ 18 years were included in the study population. This age filter was applied based on clinical relevance and data quality considerations. T1DM is rare and diagnostically complex in children under 2 years, while T2DM in individuals under 18 tends to follow different clinical trajectories and is less consistently documented in hospital EHRs. Further filters were applied to refine the study population and ensure the inclusion of informative and accurate DM cases. In this regard, only patients with at least one of the following conditions were included in the final study population (sT1DM and sT2DM): DM-specific treatment (oral antidiabetic drugs and insulins), visits to core medical departments, follow-up data, and glycated hemoglobin (HbA1c) information. These sT1DM and sT2DM were used for descriptive analysis and to calculate the prevalence of DM. Additionally, a further descriptive analysis was performed, stratifying sT1DM and sT2DM patients based on their history of hospitalization (hospitalized or attended) in the time window from 12 months before to 2 months after their inclusion.

### 2.3. Data extraction

Clinical data recorded as unstructured free text was analyzed using the EHRead® technology following previously described methods [18]. This data-driven methodology relies on NLP and ML using SNOMED CT terminology to generate a synthetic database that contains any detection of medical concepts (and associated metadata) from the source population [19–20]. Further details on the EHRead® technology including its

reading performance are provided in the Supplementary Methods and Supplementary Table 2.

### 2.4. Study variables

The full list of study variables evaluated in the unstructured data was elaborated and curated by medical experts. It included patient demographics, comorbidities, complications, laboratory parameters, and pharmacological treatments. All variables were defined as binary, except for age and laboratory parameters, which were defined as numerical. We performed a retrospective cross-sectional analysis of all individuals and variables at inclusion date —the date when individuals met all inclusion criteria— and, during the follow-up period —from inclusion to the last EHR available within the study period. To evaluate the clinical history of individuals, all available EHRs prior to the study period were considered for analyses.

Variables were analyzed using defined time windows to improve data extraction from EHRs. For repeatedly recorded variables within these windows, the value closest to the patient's inclusion date was selected. Chronic kidney disease (CKD) was defined based on clinical criteria when the following terms or clinical concepts appeared in free-text: CKD, nephrotic syndrome, impaired renal function, urine albumin $\geq$ 30 mg/24 h, urine albumin to creatinine ratio (ACR) $\geq$ 30 mg/g, urine proteins $\geq$ 150 mg/dL, creatinine $>$ 1.2 mg/dL, glomerular filtration rate (GFR) $<$ 60 mL/min/1.73 $m^2$.

### 2.5. Classification of persons with DM: Classification predictive model training, validation and data analyses

For the classification predictive model and due to the lack of consistency in reporting DM type within the EHRs, a semi-supervised ML classifier was used to assign DM type to patients with UrDM. The model examined the variable patterns of both labeled (with known DM type) and unlabeled cases (UrDM) and learned from them to classify UrDM into T1DM or T2DM. Importantly, this method allowed the incorporation of patient features beyond those that were clinically associated with DM type, using the feature patterns detected in labeled patients to inform the classification of UrDM. We defined T1DM as the positive class and T2DM as the negative class due to the lower prevalence of T1DM. This approach mitigates class imbalance and ensures that performance metrics such as precision and sensitivity provide an accurate assessment of the model's detection capabilities. The training set included all UrDM individuals and a random subset including 80% of those with reported DM type, leaving the remaining 20% for further internal validation. A total of 107 variables were evaluated for the classification (Supplementary Table 3). Sample size estimation for training the classification predictive model is provided in the Supplementary Methods section.

We defined the variables used for classification following a data-driven approach, analyzing their presence or absence in the EHRs. For example, we considered whether variables like laboratory tests were mentioned and the frequency of such mentions, rather than focusing on specific values. This approach avoided missing data by treating non-mention as zero allowing the inclusion of a wide range of potential features, despite variation in documentation frequency across concepts.

A Random Forest (RF) classifier from Scikit Learn was fitted to the training set using default hyper-parameter settings and a fixed random seed for reproducibility. Individuals with UrDM were included in training through a semi-supervised approach, combining the RF algorithm with a self-training classifier to utilize unlabeled data for improved classification and generalizability. All available data for each person was used to extract these variables, with no time point or period window restrictions.

Model performance was evaluated on the validation set using metrics such as precision, recall (sensitivity), accuracy, specificity, F1-score, and F0.5-score. We estimated 95% Confidence Intervals (CIs) through bootstrapping (1,000 splits). Additionally, we compared the semi-supervised ML classifier performance to a clinical decision algorithm based on insulin and oral hypoglycemics drugs use, age, and obesity, developed by an expert panel of two endocrinologists, a cardiologist, and a nephrologist (Supplementary Fig. 1). Given the lower prevalence of T1DM compared to that of T2DM, the evaluation prioritized approaches with higher precision and F0.5-score, aiming to minimize misclassification of T2DM as T1DM.

Hospital DM prevalence was calculated annually dividing the number of persons with DM by the number of total persons attended in the hospital that year. Crude population prevalence was calculated by dividing the number of persons with DM by the at-risk population in hospital areas, detailed by age group and sex from 1970 to 2021 Spanish demographic data. Adjusted prevalence was derived by age and sex weighting. Prevalence results for each hospital were averaged into a single estimate, and the 95% CIs were calculated.

Cumulative incidences of DM complications were estimated using the Kaplan-Meier method, excluding pre-existing conditions at inclusion and censoring individuals without events at their last follow-up. Laboratory parameter evolution was assessed with generalized additive models (GAMs), presenting prevalence, and GAM outputs with estimators and 95% CIs. Variables were summarized in descriptive tables as frequencies for categorical/binary data and means, standard deviations, or medians and quartiles for numerical data, noting the count of available values. Missing free-text information from binary variables, such as comorbidities was interpreted as no occurrence or absence of the characteristic (*i.e.*, true zero values), while multi-level categorical variables (e.g., lifestyle factors) and numeric variables (e.g., laboratory parameters) were not imputed. The frequency of available data was reported for each case.

Data analysis was performed using "R" software (version 4.0.2) and Python (version 3.7.12).

### 2.6. Ethical considerations and study approval

This study received approval from the Institutional Review Board of the reference center, Hospital Universitario La Fe in Madrid, Spain (approval number CEIm-F-PE-01–14 Ed.01-ACTA 475–12/02/2020), which also endorsed the waiver of informed consent due to the secondary use of anonymized EHR data, no direct patient interaction, minimal risk to participants, and robust data security measures to prevent re-identification. The study adheres to the Declaration of Helsinki, the General Guidance for Secondary Studies Conducted in Electronic Healthcare Databases, and the TRIPOD reporting guidelines [21]. Further details are explained in the Supplementary Methods.

## 3. Results

### 3.1. Study population

A total of 2,582,778 individuals visited the participating hospitals during the study period. From this source population, 56,181,954 EHRs were processed by EHRead®. Overall, 638,730 individuals had at least one non-negated 'DM' and/or related terms mentioned in their EHRs. Of these, 20,905 (3.3%) and 136,088 (21.3%) individuals had a reported type of T1DM and T2DM, respectively. The remaining 481,737 individuals (75.4%) had inconsistent or unregistered information in their EHRs regarding the type of DM and were labelled as UrDM. After applying the semi-supervised ML classifier and the study exclusion criteria for each DM type, 41,028 (6.5%) and 588,756 (93.5%) persons were assigned type T1DM and type T2DM, respectively (Fig. 1). The semi-supervised ML classifier outperformed the clinically guided decision algorithm with a precision of 0.956 (95% CI, 0.946 to 0.966) compared to 0.497 (95% CI, 0.484 to 0.509), and offered a more clinically plausible T2DM/T1DM ratio. Table 1 shows the performance of both models including also other metrics such as the precision,

**Table 1**
Performance comparison between the semi-supervised model and clinically guided decision algorithm.

| | Evaluation metrics | | | | | | |
|---|---|---|---|---|---|---|---|
| **Strategy** | **Specificity** | **Sensitivity** | **Precision** | **F1-score** | **F0.5-score** | **Accuracy** | **AUC-ROC** |
| **Semi-supervised ML classifier** | 0.998 (0.997, 0.998) | 0.350 (0.334, 0.364) | 0.956 (0.946, 0.966) | 0.512 (0.494, 0.527) | 0.710 (0.695, 0.723) | 0.913 (0.911, 0.915) | 0.805 (0.796, 0.814) |
| **Clinical decision algorithm** | 0.925 (0.922, 0.928) | 0.497 (0.481, 0.512) | 0.497 (0.484, 0.51) | 0.497 (0.484, 0.509) | 0.497 (0.485, 0.509) | 0.869 (0.866, 0.872) | NA* |

| | Diabetes mellitus population groups | | |
|---|---|---|---|
| | **T1DM** | **T2DM** | **T2DM/T1DM ratio** |
| **Population by semi-supervised ML classifier** | 41,028 (6.5) | 588,756 (93.5) | 14.3 |
| **Population by clinical decision algorithm** | 150,741 (23.9) | 478,921 (76.1) | 3.2 |

Abbreviations: T1DM, type 1 diabetes mellitus; T2MD, type 2 diabetes mellitus; AUC-ROC, area under the curve of the receiver operating characteristic; NA, not applicable. Results for the validation set only. Findings are reported as mean and 95% confidence interval. Metrics indicate poor ($\leq 0.5$), modest (0.5–0.7), moderate (0.7–0.85) or high ($\geq 0.85$) predictive performance.

* The clinically guided decision algorithm is rule-based and does not output probabilities, then AUC-ROC could not be computed.

sensitivity, F1-score, F0.5-score, and accuracy as well as the final populations obtained. After excluding the possible false positive and non-informative individuals, a total of 30,441 individuals with T1DM and 448,733 individuals with T2DM comprised the final subpopulations sT1DM and sT2DM, respectively (Fig. 1). Patient characteristics with respect to the variables used to define the subpopulations are shown in Supplementary Table 4.

### 3.2. Prevalence of T1DM and T2DM in the hospital setting

Prevalence calculations by year and by DM type subpopulation, including subgroup analysis stratified by age group and sex are shown in Table 2. Overall, DM prevalence ranged between 28.3%-30.7% in the study period with prevalence in sT1DM and sT2DM ranging between 1.9%-2.1% and 26.2%-28.6%, respectively. Overall, the crude prevalence was 15.1% for both males and females, with rates of 0.5–0.6% for T1DM and 14.7–14.5% for T2DM, respectively for each sex. The crude prevalence of T1DM remained relatively stable across different age groups and sexes. However, T2DM prevalence showed a progressive increase related to age: it was recorded at 11.6% (95% CI, 7.4 to 15.8) in individuals aged 45–59 years, increased to 19.1% (95% CI, 12.9 to 25.3) in those aged 60–74 years, and reached 28.5% (95% CI, 18.0 to 39.0) in the group over 75 years (Table 2). The age-adjusted prevalence of DM is reported in Supplementary Table 5.

### 3.3. Demographic and clinical characteristics of persons with DM in the hospital setting

Demographic, clinical characteristics, and laboratory parameters were described in the sT1DM and sT2DM subpopulations (Table 3). For both subpopulations the proportion of men and women was close to 50% (women, 53.1% and 52.3%, respectively). The median (Q1; Q3) age of sT1DM individuals was 17 years (11; 45), of whom 12.1% were older than 65 years at inclusion. The most common registered comorbid conditions in both sT1DM and sT2DM individuals were hypertension (32.9% and 62.1%, respectively), dyslipidemia (29.8% and 58.1%), CKD (25.7% and 28.5%), ischemic heart disease (IHD, 10.6% and 18.8%), and chronic heart failure (CHF, 12.6% and 16.9%). For all parameters, the percentage of available information in the free text of EHRs was less than 50%. Among sT1DM individuals with available data, main values (median [Q1; Q3]) included: HbA1c, 7.3% (5.8; 8.7), with 56.1% of persons having HbA1c levels $\geq$ 7%; serum creatinine, 0.7 mg/dL (0.5; 0.9); estimated glomerular filtration rate (eGFR), 99.8 mL/min/1.73 $m^2$ (72.3; 129.6), with 14.5% of individuals having an eGFR $<$ 60 mL/min/1.73 $m^2$; low density lipoprotein (LDL) cholesterol, 99 mg/dL (78; 122), with 51.2% of individuals having LDL levels $<$ 100 mg/dL. Among sT2DM individuals, the corresponding values were: HbA1c, 6.2% (5.6; 7.1), with 29.5% of persons having HbA1c levels $\geq$ 7%; serum creatinine, 0.9 mg/dL (0.7; 1.1); eGFR, 78.4 mL/min/1.73 $m^2$ (58.9; 97.6), with 26.1% of individuals having an eGFR $<$ 60 mL/min/1.73 $m^2$; LDL cholesterol, 99 mg/dL (75; 126), with 50.4% of individuals having LDL levels $<$ 100 mg/dL. Information on albuminuria was available for 6.5% of sT1DM, of whom 75.8% had microalbuminuria and 1.5% had macroalbuminuria, and for 3.1% of sT2DM, of whom 51.9% had microalbuminuria and 2.3% had macroalbuminuria. For the sT1DM and sT2DM subpopulations, DM-specific treatments as well as other pharmacological therapies for the management of comorbidities and complications associated with DM are shown in Supplementary Table 6, including antiplatelet agents (10.6% in sT1DM and 25.6% in sT2DM), anticoagulants (3.3% and 11.5%), antihypertensives (11.8% and 23.7%), and statins (6.4% and 14.6%).

The analysis of clinical characteristics and laboratory parameters related to health metrics in persons with T1DM and T2DM (some of them critical for assessing cardiovascular risk) revealed a notable proportion of missing data in both groups (Table 3). For example, HbA1c data was available only for 12.2% of T1DM and 10.7% of T2DM persons. Similarly, other key clinical and laboratory variables showed high proportions of missing data across both T1DM and T2DM groups, with availability often below 15%.

When analyzing attended and hospitalized patients, we observed similar demographic characteristics among individuals with T1DM. In contrast, among those with T2DM, the median age (Q1, Q3) was 67 years (52, 79) for hospitalized patients and 57 years (42, 71) for attended patients. In both T1DM and T2DM populations, hospitalized patients more frequently reported smoking history and relevant comorbidities such as CKD, IHD, or CHF, and notable differences in laboratory parameters were observed between attended and hospitalized patients. Detailed information on these data is provided in Supplementary Table 7.

### 3.4. Evolution of DM individuals in the hospital setting

We followed up 88.0% of the sT1DM and 85.5% of the sT2DM subpopulations for a minimum of 12 months, and 37.6% and 35.2% (respectively) for at least 60 months. The median (Q1; Q3) follow-up for sT1DM was 40.1 months (23.2; 55.2), and 38.2 months (20.3; 54.0) for sT2DM. Analysis of the cumulative incidence of DM-related complications is shown in Fig. 2. In both diagnostic groups, CKD, CHF and IHD were the DM-related complications with the highest cumulative incidence during follow-up. In sT1DM, the cumulative incidence of these complications at 60 months was 20.8%, 16.0% and 11.7%, respectively. In sT2DM individuals, the corresponding incidences were 28.2%, 22.1%

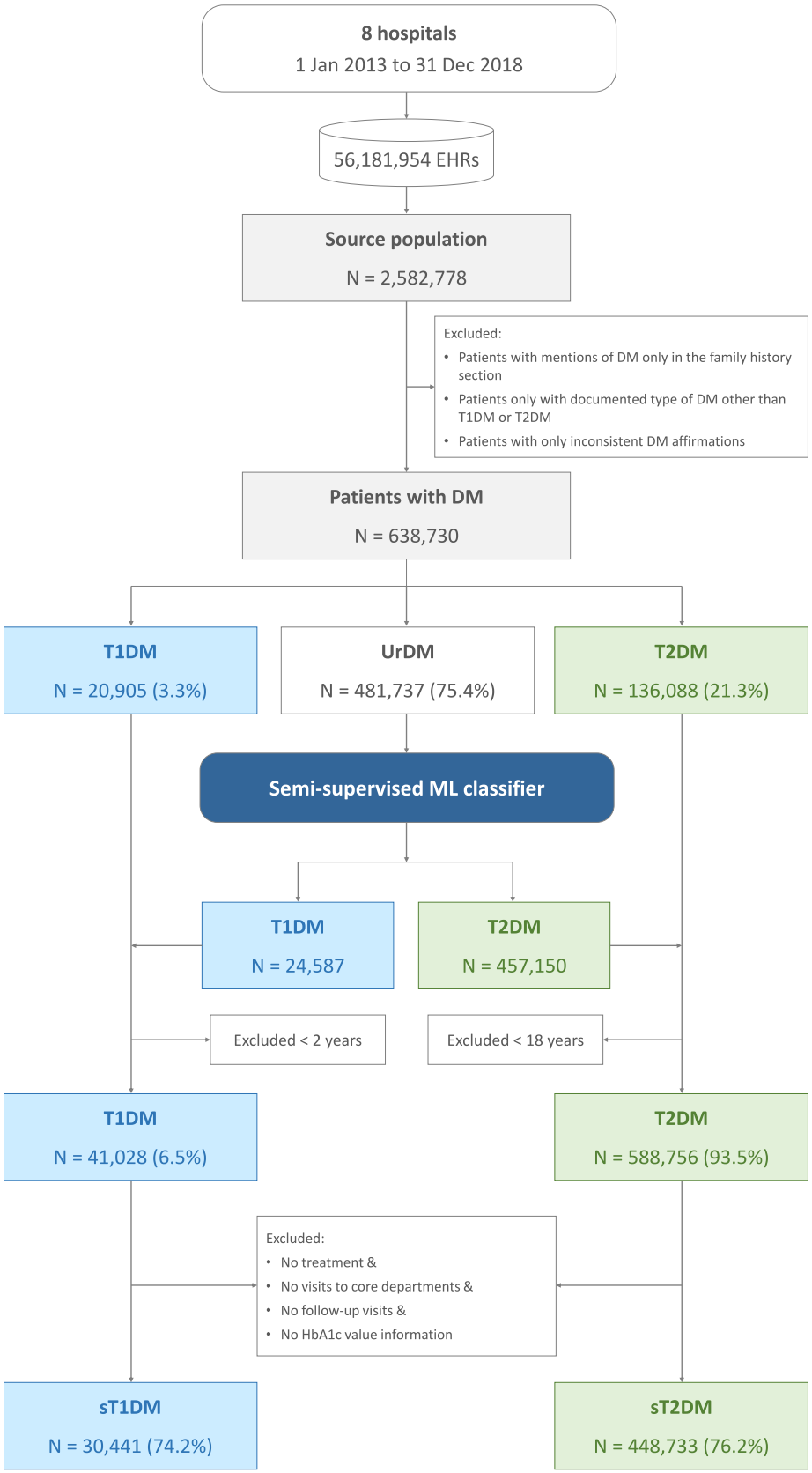


**Fig. 1.** Patient flow chart. Abbreviations: DM, diabetes mellitus; EHR, electronic health record; HbA1c, glycated hemoglobin; ML, machine learning; sT1DM, T1DM subpopulation which excluded non-informative individuals; sT2DM, T2DM subpopulation which excluded non-informative individuals; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; UrDM, unregistered type of diabetes mellitus.

**Table 2**
Prevalence of diabetes.

**Attended population prevalence, % (95% CI)***

| Year | Overall | T1DM | T2DM |
|---|---|---|---|
| **2013**[a] | 28.9 (23.4–34.4) | 1.9 (1.2–2.6) | 27.0 (21.4–32.6) |
| **2014** | 30.4 (26.8–34.0) | 2.1 (1.4–2.8) | 28.3 (24.8–31.8) |
| **2015** | 30.7 (27.1–34.3) | 2.1 (1.4–2.8) | 28.6 (25.1–32.0) |
| **2016** | 30.5 (26.9–34.1) | 2.1 (1.4–2.7) | 28.4 (25.0–31.8) |
| **2017** | 30.1 (26.5–33.7) | 2.0 (1.3–2.6) | 28.1 (24.7–31.6) |
| **2018**[a] | 28.3 (23.8–32.7) | 2.1 (1.3–2.8) | 26.2 (22.1–30.3) |

**Crude prevalence, % (95%CI)**

| Age group | Overall | | | T1DM | | | T2DM | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Male | Female | All | Male | Female | All | Male | Female |
| **All** | 15.1 (9.3–20.9) | 15.1 (9.4–20.9) | 15.1 (9.2–20.9) | 0.5 (0.3–0.7) | 0.5 (0.3–0.6) | 0.6 (0.4–0.8) | 14.6 (9.0–20.2) | 14.7 (9.1–20.3) | 14.5 (8.8–20.1) |
| **18–29 years** | 10.1 (5.7–14.5) | 9.1 (5.1–13.1) | 11.1 (6.3–16.0) | 1.0 (0.4–1.5) | 0.8 (0.3–1.3) | 1.1 (0.5–1.7) | 9.1 (5.2–13.0) | 8.3 (4.7–11.8) | 10.0 (5.8–14.3) |
| **30–44 years** | 10.8 (6.0–15.6) | 10.1 (5.6–14.5) | 11.5 (6.4–16.7) | 0.6 (0.4–0.9) | 0.4 (0.3–0.6) | 0.8 (0.5–1.2) | 10.2 (5.6–14.8) | 8.3 (4.7–11.8) | 10.0 (5.8–14.3) |
| **45–59 years** | 11.9 (7.6–16.3) | 12.2 (7.9–16.5) | 11.7 (7.4–16.0) | 0.4 (0.3–0.5) | 0.4 (0.2–0.5) | 0.4 (0.3–0.6) | 11.6 (7.4–15.8) | 11.8 (7.6–16.1) | 11.3 (7.1–15.5) |
| **60–74 years** | 19.5 (13.2–25.8) | 21.8 (14.9–28.6) | 17.5 (11.6–23.4) | 0.4 (0.3–0.6) | 0.4 (0.3–0.6) | 0.4 (0.3–0.6) | 19.1 (12.9–25.3) | 21.3 (14.7–28.0) | 17.1 (11.4–22.9) |
| **> 75 years** | 29.0 (18.3–39.7) | 31.5 (20.1–42.8) | 27.3 (17.0–37.5) | 0.5 (0.3–0.7) | 0.5 (0.3–0.7) | 0.5 (0.3–0.7) | 28.5 (18.0–39.0) | 30.9 (19.8–42.1) | 26.8 (16.8–36.8) |

Abbreviations: CI, confidence interval; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.
* Results were aggregated by hospital and presented as mean prevalence (95% CI) for each year of the study period.
[a] Three of the eight participating hospitals had only partial data for 2013 and 2018 and were therefore excluded from the analysis for these years. Attended population includes hospitalized and ambulatory patients attended to the participating hospitals.

and 17.0%, respectively. All other DM-related complications, regardless of their relative incidence, also increased 2- to 3-fold during the study period in both sT1DM and sT2DM subpopulations (Supplementary Table 8).

Smooth plots showing the values of the main extracted laboratory parameters recorded in the unstructured data over the follow-up period are shown in Supplementary Fig. 2. Overall, data on laboratory parameters were reported for fewer than 20% of the included persons. HbA1c levels exhibited a slight decrease in T1DM and stability in T2DM, GFR and LDL cholesterol levels decreased, whereas serum creatinine and high-density lipoprotein (HDL) cholesterol levels increased over time in both sT1DM and sT2DM individuals.

## 4. Discussion

Unstructured medical data, such as clinical notes, often captures richer and more nuanced information than structured data, providing narrative and contextual details that predefined forms cannot [22]. However, analyzing and exploiting unstructured data requires advanced AI techniques like NLP and ML. To the extent of our knowledge, this work represents the most significant attempt to leverage unstructured data recorded in patients' EHRs in the field of DM in Spain. Using these technologies, our study highlights the high prevalence of DM in Spanish hospitals, with a notable burden of cardiovascular and renal comorbidities and complications. Although our findings demonstrate the potential of these technologies to address this challenge, they also indicate room for improvement in leveraging the information contained within EHR narratives.

First, our data highlight the need for enhanced documentation practices emphasizing the importance and necessity of generating comprehensive reports in EHR narratives in the clinical environment. Specifically, we found that only about a quarter of the subjects in our cohort of individuals with DM had registered their specific DM type in line with previous publications showing that half of DM individuals had unregistered type of DM [9,11]. Despite this, AI techniques allowed us to develop a semi-supervised ML classifier with better classification metrics and more clinically plausible T2DM/T1DM ratio compared to a clinically guided algorithm, thereby enabling prevalence calculations. The results showed that the prevalence of T1DM remained stable throughout the study period (ranging from 1.9% to 2.1%) whereas T2DM prevalence showed a more noticeable gradual shift (from 26.2% to 28.6%). In this regard, persons with T2DM are often hospitalized due to conditions other than DM [23–26], and detailed information on their clinical profiles in hospital settings is crucial for devising strategies to enhance disease management and minimize the risk of complications. Interestingly, we found that among included T2DM persons, those who required hospitalization were older, more likely to smoke, and had more comorbidities than those who were attended without the need for hospitalization.

Approximately 30% of individuals with T1DM and 60% of those with T2DM in our cohort had documented hypertension, aligning with previous reports of 30–40% for T1DM [27], and up to 80% for T2DM [28]. Dyslipidemia, a key cardiovascular risk factor [30], was reported in about a third of individuals with T1DM and over a half of T2DM individuals, with literature rates ranging from 28% to 67% in T1DM [29] and 25% to 95% in T2DM [30–31], depending on population and definitions used, complicating direct comparisons with our findings.

CHF and CKD, frequently co-occurring in DM, are mutually exacerbating conditions that increase risks of hospitalization, intensive care, kidney replacement therapy, and mortality [32–42]. In our cohort, at inclusion, IHD or CHF was reported in 10% of individuals with T1DM and 18% of those with T2DM, while CKD was noted in 25% of T1DM and 28% of T2DM cases. Population registries report lower rates of IHD (4.2% in T1DM, 8.6% in T2DM) and CHF (3.1% in T1DM, 7.5–15% in T2DM) [41]. CKD prevalence in T1DM aligns with previous estimates (18–20%) [35,38,43], while T2DM shows a broader range (21–50%) influenced by diagnostic criteria, demographics, and medical management [34–35,38,43–44]. Notably, a prior Spanish study in primary care practices reported a 24% CKD prevalence in T2DM [43], consistent with our findings.

The higher frequencies observed in our study may be related with the characteristics of our source population, which include individuals

**Table 3**
Patient demographic, clinical and analytical characteristics at baseline.

| Characteristic | sT1DM (N = 30,441) | sT2DM (N = 448,733) |
|---|---|---|
| **Demographics and toxic habits** | | |
| **Female sex, n (%)**[a] | 16,160 (53.1) | 234,544 (52.3) |
| **Age, years**[a] | | |
| Median (Q1, Q3) | 17 (11, 45) | 61 (46, 75) |
| ≥ 65 years | 3,692 (12.1) | 198,386 (44.2) |
| **BMI, kg/m$^2$** | | |
| Available, n (%) | 2673 (8.8) | 18,456 (4.1) |
| Median (Q1,Q3) | 21.1 (16.8, 26.5) | 29 (24.3, 34.2) |
| **Smoking history**[b] | | |
| Available, n (%) | 6,462 (21.2) | 164,245 (36.6) |
| Current smoker, n (%) | 2,444 (37.8) | 60,672 (36.9) |
| Ex-smoker, n (%) | 1,495 (23.1) | 63,207 (38.5) |
| Never smoker, n (%) | 2,523 (39.1) | 40,366 (24.6) |
| **Comorbidities and diabetes-related complications, n (%)**[c] | | |
| Hypertension | 10,009 (32.9) | 278,534 (62.1) |
| Dyslipidemia | 9,081 (29.8) | 260,602 (58.1) |
| Chronic kidney disease | 7,831 (25.7) | 127,872 (28.5) |
| Ischemic heart disease | 3,225 (10.6) | 84,438 (18.8) |
| Chronic heart failure | 3,823 (12.6) | 75,808 (16.9) |
| Peripheral vascular disease | 2,030 (6.7) | 58,765 (13.1) |
| Atrial fibrillation | 2,089 (6.9) | 53,661 (12.0) |
| Obesity* | 1,729 (5.7) | 40,960 (9.1) |
| Cerebrovascular disease | 1,288 (4.2) | 33,384 (7.4) |
| Peripheral arterial disease | 442 (1.5) | 13,756 (3.1) |
| Diabetic retinopathy | 1,461 (4.8) | 10,470 (2.3) |
| Diabetic neuropathy | 1,051 (3.5) | 4,772 (1.1) |
| Foot amputation | 64 (0.2) | 1,123 (0.3) |
| **Laboratory parameters**[d] | | |
| **HbA1c [%]**[i] | | |
| Available, n (%) | 3,700 (12.2) | 48,211 (10.7) |
| Median (Q1,Q3) | 7.3 (5.8, 8.7) | 6.2 (5.6, 7.1) |
| HbA1c ≥ 7%, n (%)[¥] | 2,077 (56.1) | 14,246 (29.5) |
| **Serum creatinine [mg/dL]**[i] | | |
| Available, n (%) | 11,056 (36.3) | 221,503 (49.4) |
| Median (IQR) | 0.7 (0.5–0.9) | 0.9 (0.7–1.1) |
| Creatinine ≥ 1.2 mg/dL, n (%) | 1,145 (10.4) | 43,761 (19.8) |
| **GFR [mL/min/1.73 m$^2$]** ** and [i] | | |
| Available, n (%) | 9,364 (30.8) | 222,969 (49.7) |
| Median (Q1, Q3) | 99.8 (72.3, 129.6) | 78.4 (58.9, 97.6) |
| GFR ≥ 90[¥], n (%) | 5,619 (60.0) | 76,772 (34.4) |
| GFR 60–90[¥], n (%) | 2,391 (25.5) | 88,009 (39.5) |
| GFR 30–60[¥], n (%) | 972 (10.4) | 45,565 (20.4) |
| GFR < 30[¥], n (%) | 382 (4.1) | 12,623 (5.7) |
| **Albuminuria, n (%)**[¥],*** | | |
| Available, n (%) | 1,970 (6.5) | 14,068 (3.1) |
| Normoalbuminuria | 447 (22.7) | 6,452 (45.9) |
| Microalbuminuria | 1,494 (75.8) | 7,295 (51.9) |
| Macroalbuminuria | 29 (1.5) | 321 (2.3) |
| **Lipid profile** | | |
| Total cholesterol [mg/dL][i] | | |
| Available, n (%) | 5,833 (19.2) | 90,629 (20.2) |
| Median (Q1, Q3) | 169 (143, 201) | 170 (138, 205) |
| Total cholesterol < 200 mg/dL[¥], n (%) | 4,287 (73.5) | 64,934 (71.6) |
| HDL [mg/dL][i] | | |
| Available, n (%) | 3,616 (11.9) | 63,320 (14.1) |
| Median (Q1, Q3) | 51 (40, 63) | 45 (35, 56) |
| HDL > 50 mg/dL[¥], n (%) | 1,968 (54.4) | 24,183 (38.2) |
| LDL [mg/dL][i] | | |
| Available, n (%) | 4,302 (14.1) | 70,870 (15.8) |
| Median (Q1, Q3) | 99 (78, 122) | 99 (75, 126) |
| LDL < 100 mg/dL[¥], n (%) | 2,204 (51.2) | 35,697 (50.4) |
| LDL < 70 mg/dL[¥], n (%) | 720 (16.7) | 13,874 (19.6) |
| Triglycerides [mg/dL] | | |
| Available, n (%) | 5,631 (18.5) | 87,961 (19.6) |
| Median (Q1, Q3) | 89 (63, 134) | 112 (81, 161) |

Abbreviations: BMI, body mass index; GFR, glomerular filtration rate; HbA1c, glycated hemoglobin; HDL, high density lipoprotein; LDL: low density lipoprotein; IQR, interquartile range; SD, standard deviation; sT1DM, final type 1 diabetes mellitus subpopulation; sT2DM, final type 2 diabetes mellitus subpopulation.

[a] All calculated percentages include patients with missing information.
[b] Data extracted and analyzed considering all information up to 2 months after inclusion date.
[c] Data extracted and analyzed considering all information up to the inclusion date.
[d] Data extracted and analyzed considering all information between 12 months before index date and up to 2 months after index date.
[*] Obesity was extracted as both direct variable (e.g., mention of obesity in EHRs) and inferred variable (i.e., BMI ≥ 30).
[**] GFR data were extracted as both direct variable (i.e., mention of GFR value in EHRs) and inferred variable (i.e., a posteriori calculation using the CKD-EPI creatinine equation, with available patient information, including sex, age, and creatinine).
[***] Albuminuria data were extracted from detection of the terms "normo/micro/macroalbuminuria" in the free text.
[¥] Percentage of patients with at least one laboratory value meeting the condition in the analyzed time window; data calculated considering only patients with available information on the specified laboratory parameter.
[i] For analytical parameters, if multiple datapoints were available for a single patient, the value closest to the end of the year was considered for analysis.

managed in hospital settings, with higher incidence of general and cardiovascular comorbidities. However, it is worth noting that although hospitalized T2DM could have more complex clinical profiles, potentially due to a higher burden of comorbidities, this observation may also be influenced by reporting bias, as hospitalized individuals often have more detailed documentation of their medical history. This could limit direct comparisons with other population-based studies.

Regarding treatment for comorbidities, acetylsalicylic acid was the most used antiplatelet, acenocoumarol the main anticoagulant, and statins the most common lipid-lowering therapy, with first-line agents dominating antihypertensive use. However, treatments were underreported, potentially underestimating their actual use in this population.

On the other hand, our study observed a 2- to 3-fold increase in the cumulative incidence of DM-related complications, including CHF, IHD, and CKD, during the 60-month follow-up period across all individuals. These findings underscore the high-risk profile of this population and highlight the need of close monitoring to manage diabetes and address modifiable factors contributing to CKD and cardiovascular complications [45].

Despite a large cohort size, a significant proportion of data was missing for several key health metrics, including HbA1c, BMI, albuminuria, and LDL cholesterol levels. For instance, over 85% of the data for most of these variables was unavailable, which limits the ability to draw comprehensive conclusions about the entire patient population. In the context of a descriptive study with high percentages of missing data, we did not impute these missing values to avoid potential inaccuracies [46]. Additionally, in this case, the missing data did not occur completely at random, as there is evidence that in EHRs, altered data are reported more frequently [47]. Therefore, imputation could have introduced a systematic bias that would undermine the validity of any imputed values. It is worth noting that these findings related to missing data underscore the necessity for a combined analysis of structured and unstructured data to assess these variables more effectively. This can provide a more comprehensive and accurate understanding of patients' clinical conditions enhancing the quality and depth of insights derived from EHR data, ultimately improving patient care and clinical decision-making [48–50].

The main strength of this multicenter study is the analysis of a vast amount of EHRs from 8 Spanish hospitals spanning a study period of 6 years. The application of EHRead® NLP technology allowed us to extract and interpret unstructured RWD to estimate the prevalence of DM, T1DM and T2DM, in the Spanish population attending the hospital setting. On the other hand, the superior performance of our model compared to the clinical decision algorithm reflects the capability of ML to uncover hidden patterns in the data, extending beyond traditionally known variables. However, there are also some limitations, mainly tied

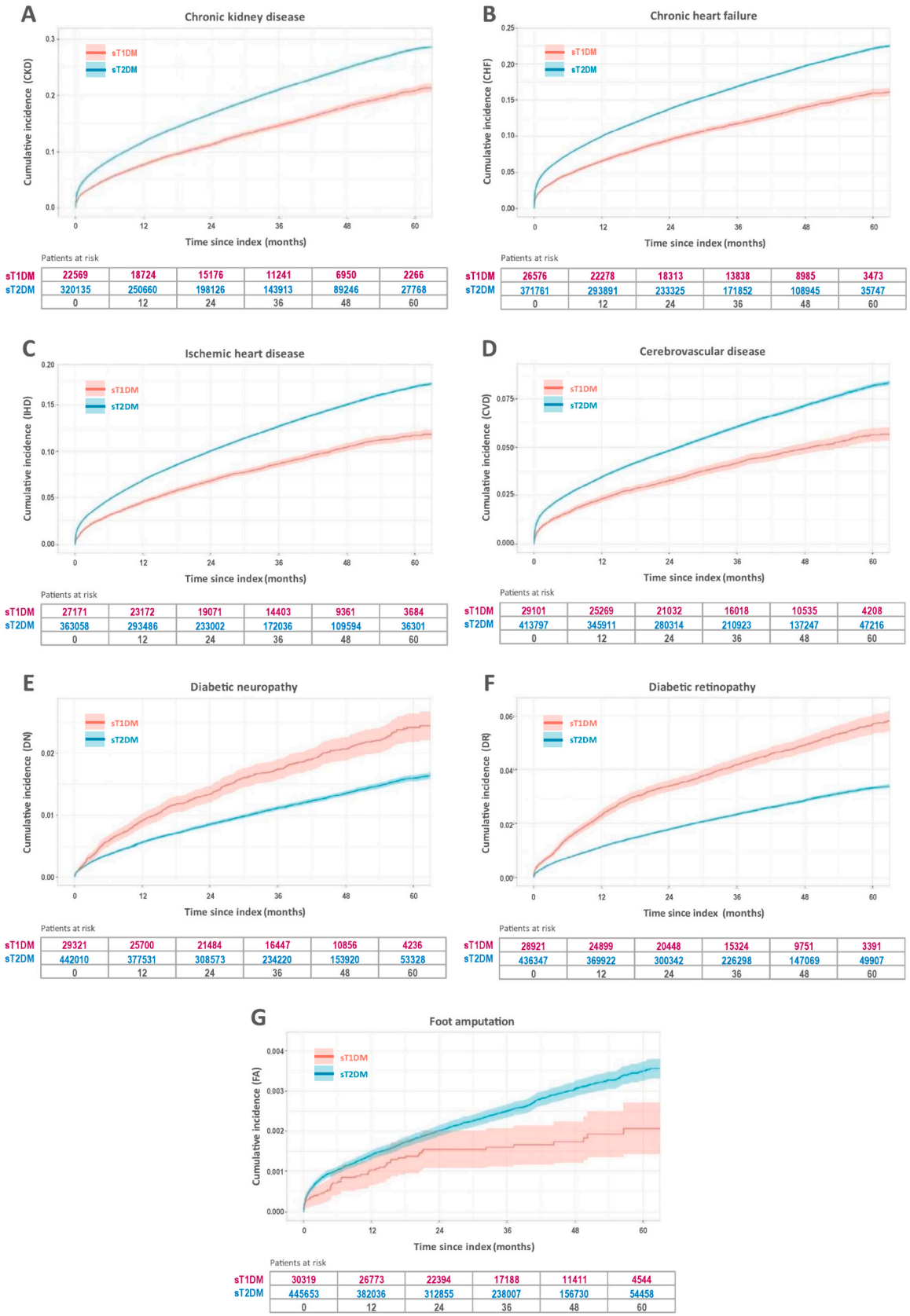


**Fig. 2.** Cumulative incidence of comorbidities and diabetes-related complications during the study period. Abbreviations: sT1DM, type 1 diabetes mellitus subpopulation; sT2DM, type 2 diabetes mellitus subpopulation.

to data source and extraction challenges. Some important variables to characterize DM are more frequently reported as structured information within the EHRs. Then, in our study, the lack of comprehensive documentation on critical variables like laboratory results or concurrent treatments, restricts deriving conclusive insights, limiting the detailed characterization of the clinical phenotype of the study population. In this regard, the inclusion of other important variables, such as vital signs —including blood pressure values— would have been of interest. Unfortunately, those parameters were less present than laboratory variables in free text with much smaller number of recordings per person, and thus, not available to include as a result. Additionally, focusing on the population attended in the hospital setting, which presents high comorbidity and complexity, may restrict the generalizability of the results to a broader population. Moreover, other ensemble methods could be explored to further improve the performance of semi-supervised model to classify individuals with unregistered type of DM.

In summary, this is the first study using unstructured data extracted with NLP and ML techniques to assess the prevalence and clinical characteristics of DM individuals attending tertiary hospitals in Spain, including both hospitalized and ambulatory individuals. Our findings showed a high prevalence of the disease in the study population with a significant burden of comorbidities and complications, particularly CKD and CHF. While information extracted from free-text narratives provides a valuable resource for improving the understanding of this condition, it is crucial to improve data recording practices to maximize the utility of NLP technology. Additionally, future studies should prioritize integrating unstructured data with structured data to overcome current limitations. This integration would enable a more comprehensive and accurate analysis of patient information, ultimately enhancing informed clinical decision-making and healthcare strategies.

Beyond its capacity for massive data extraction reported in EHRs, the use of NLP and ML techniques holds considerable promise for real-time clinical decision-making. These technologies can facilitate the rapid identification of high-risk individuals and detect early signs of complications, such as CKD and CHF, directly from unstructured clinical narratives. These tools can enable timely interventions, optimize resources, and personalize treatment plans. Further studies are needed to develop algorithms that can reliably perform these identifications advancing proactive, data-driven diabetes management and improving patient outcomes.

## Conflict of interest

A.S.JJ, S.M.M, P.B, and P.D.A are AstraZeneca employees.

A-C.C and F.V are Medsavana and Savana Research SL employees.

B-C.AJ, P.dI.L, N.G.JF, and M−T.JF are members of the external scientific committee of the Diabetic@ study that has been sponsored and funded by AstraZeneca.

N.GJF acts as external advisor and expert speaker for AbbVie, Amgen, AstraZeneca, Bayer, Boehringer Ingelheim, Esteve, Genzyme, Lilly, MSD, Novartis, NovoNordisk, Servier, Shire and Vifor Pharma.

## CRediT authorship contribution statement

**A.J. Blanco-Carrasco:** Writing – original draft, Formal analysis, Data curation, Conceptualization. **J.F. Merino-Torres:** Writing – original draft, Formal analysis, Conceptualization. **M. Rubio Almanza:** Writing – review & editing. **G. Canovas Molina:** Writing – review & editing. **M.A. Brito-Sanfiel:** Writing – review & editing. **D.E. Barajas Galindo:** Writing – review & editing. **L.A. Cuellar Olmedo:** Writing – review & editing. **D. Mauricio:** Writing – review & editing. **S. Tofé Povedano:** Writing – review & editing. **J.A. Balsa Barro:** Writing – review & editing. **J.J. Aparicio Sánchez:** Writing – review & editing, Supervision, Funding acquisition. **M. Sequera Mutiozabal:** Writing – review & editing, Project administration, Funding acquisition. **B. Pimentel:** Writing – review & editing, Project administration, Funding acquisition. **A. Pérez Domínguez:** Writing – review & editing, Project administration, Funding acquisition. **C. Arias-Cabrales:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **V. Fanjul:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **L. Pérez de Isla:** Writing – review & editing, Writing – original draft. **J.F. Navarro González:** Writing – review & editing, Writing – original draft.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Savana Research Group members include Sebastian Menke, Ignacio Salcedo, Natalia Polo, Judith Marín-Corral, David Casadevall, and Miren Taberna. We also thank Luis Fernández for his contributions.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.diabres.2025.112214.

## References

[1] Diabetes [article online], 2023. Available from https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 20 April 2023.
[2] Standl E, Khunti K, Hansen TB, Schnell O. The global epidemics of diabetes in the 21st century: Current situation and perspectives. Eur J Prev Cardiol 2019;26:7–14.
[3] IDF Diabetes Atlas - 10th Edition [article online], 2021. Available from https://diabetesatlas.org/data/en/world/. Accessed 26 April 2023.
[4] Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022;183:109119.
[5] Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. Diabetologia 2019;62:3–16.
[6] Martínez LC, Zahra T. Chronic Complications of Diabetes. EMJ Diabet 2022.
[7] Kufeldt J, Kovarova M, Adolph M, Staiger H, Bamberg M, Häring HU, et al. Peter A. Prevalence and Distribution of Diabetes Mellitus in a Maximum Care Hospital: Urgent Need for HbA1c-Screening. Experimental and clinical endocrinology & diabetes : official journal, German Society of Endocrinology [and] German Diabetes Association 2018;126:123-129.
[8] Tönnies T, Rathmann W, Hoyer A, Brinks R, Kuss O. Quantifying the underestimation of projected global diabetes prevalence by the International Diabetes Federation (IDF) Diabetes Atlas. BMJ open diabetes research & care 2021;9.
[9] American Diabetes Association Professional Practice C. Summary of Revisions: Standards of Care in Diabetes-2025. Diabetes Care 2025;48:S6–S13.
[10] Stone MA, Camosso-Stefinovic J, Wilkinson J, de Lusignan S, Hattersley AT, Khunti K. Incorrect and incomplete coding and classification of diabetes: a systematic review. Diabetic Medicine : A Journal of the British Diabetic Association 2010;27:491–7.
[11] de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabetic Medicine : A Journal of the British Diabetic Association 2012;29:181–9.
[12] de Lusignan S, Crawford L, Munro N. Creating and using real-world evidence to answer questions about clinical effectiveness. J Innov Health Inform 2015;22:368–73.
[13] de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabetic Medicine : A Journal of the British Diabetic Association 2010;27:203–9.
[14] Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform 2018;88:11–9.
[15] Cosin-Sales J, Anguita M, Suarez C, Arias-Cabrales C, Martinez-Sanchez L, Savana Research G, et al. Oral anticoagulant treatment in atrial fibrillation: the AFIRMA real-world study using natural language processing and machine learning. Rev Clin Esp (Barc) 2024;224:428–36.
[16] Roman Ivorra JA, Trallero-Araguas E, Lopez Lasanta M, Cebrian L, Lojo L, Lopez-Muniz B, et al. Prevalence and clinical characteristics of patients with rheumatoid arthritis with interstitial lung disease using unstructured healthcare data and machine learning. RMD Open 2024;10.

[17] Turchin A, Florez Builes LF. Using Natural Language Processing to Measure and Improve Quality of Diabetes Care: A Systematic Review. J Diabetes Sci Technol 2021;15:553–60.
[18] Izquierdo JL, Morena D, González Y, Paredero JM, Pérez B, Graziani D, et al. Clinical Management of COPD in a Real-World Setting. A Big Data Analysis Archivos de bronconeumologia 2021;57:94–100.
[19] Benson T. Principles of Health Interoperability HL7 and SNOMED. Springer, Ed. London, UK, 2012.
[20] Espinosa-Anke L, Tello J, Pardo A, Medrano I, Ureña A, Salcedo I, et al. Savana: A global information extraction and terminology expansion framework in the medical domain. Procesamiento del Lenguaje Natural 2016;57:23–30.
[21] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162: W1–.
[22] Agrawal S, George R, Prabhakar Vaidya V, Chakkrapani S, Prajapati R, et al. Development of natural language processing (NLP) models for extracting key features from unstructured notes to create real-world data (RWD) assets for clinical research at scale. In *2023 ASCO Annual Meeting*, Journal of Clinical Oncology, 2023.
[23] Tinajero MG, Malik VS. An Update on the Epidemiology of Type 2 Diabetes: A Global Perspective. Endocrinol Metab Clin North Am 2021;50:337–55.
[24] Fox CS, Coady S, Sorlie PD, D'Agostino Sr RB, Pencina MJ, Vasan RS, et al. Increasing cardiovascular disease burden due to diabetes mellitus: the Framingham Heart Study. Circulation 2007;115:1544–50.
[25] Schütt M, Fach EM, Seufert J, Kerner W, Lang W, Zeyfang A, et al. Diabetic Medicine : A Journal of the British Diabetic Association 2012;29:e176–9.
[26] Zimmet PZ, Magliano DJ, Herman WH, Shaw JE. Lancet Diabetes Endocrinol 2014; 2:56–64.
[27] Maahs DM, Kinney GL, Wadwa P, Snell-Bergeon JK, Dabelea D, Hokanson J, et al. Ambulatory blood pressure and arterial stiffness in individuals with type 1 diabetes. Diabetologia 2018;61:1935–45.
[28] Pantalone K.M., Hobbs T.M., Wells B.J., Kong S.X., Kattan M.W., Bouchard J., Yu C., Sakurada B., Milinovich A., Weng W., Bauman J.M., Zimmerman R.S. Clinical characteristics, complications, comorbidities and treatment patterns among patients with type 2 diabetes mellitus in a large integrated health system. BMJ Open Diabetes Res Care 2015; Jul 22;3(1):e000093.
[29] Pérez A, Wägner AM, Carreras G, Giménez G, Sánchez-Quesada JL, Rigla M, et al. Prevalence and phenotypic distribution of dyslipidemia in type 1 diabetes mellitus: effect of glycemic control. Arch Intern Med 2000;160:2756–62.
[30] Poznyak A, Grechko AV, Poggio P, Myasoedova VA, Alfieri V, Orekhov AN. The diabetes mellitus-atherosclerosis connection: the role of lipid and glucose metabolism and chronic inflammation. International Journal of Molecular Sciences 2020;21.
[31] Chehade JM, Gladysz M, Mooradian AD. Dyslipidemia in type 2 diabetes: prevalence, pathophysiology, and management. Drugs 2013;73:327–39.
[32] House AA, Wanner C, Sarnak MJ, Pina IL, McIntyre CW, Komenda P, et al. Heart failure in chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int 2019;95:1304–17.
[33] Khunti K, Charbonnel B, Chen H, Cherney DZ, Cooper A, Fenici P, et al. Prevalence and progression of chronic kidney disease among patients with type 2 diabetes: Insights from the DISCOVER study. Diabetes Obes Metab 2021;23:1956–60.
[34] Majeed MS, Ahmed F, Teeling M. The Prevalence of Chronic Kidney Disease and Albuminuria in Patients With Type 1 and Type 2 Diabetes Attending a Single Centre. Cureus 2022;14:e32248.
[35] Martínez Candela J, Sangrós González J, García Soidán FJ, Millaruelo Trillo JM, Díez Espino J, Bordonaba Bosque D, et al. Chronic renal disease in spain: prevalence and related factors in persons with diabetes mellitus older than 64 years. Nefrologia 2018;38:401–13.
[36] Mata-Cases M, Franch-Nadal J, Real J, Vlacho B, Gómez-García A, Mauricio D. Evaluation of clinical and antidiabetic treatment characteristics of different subgroups of patients with type 2 diabetes: Data from a Mediterranean population database. Prim Care Diabetes 2021;15:588–95.
[37] Ohta M, Babazono T, Uchigata Y, Iwamoto Y. Comparison of the prevalence of chronic kidney disease in Japanese patients with Type 1 and Type 2 diabetes. Diabetic Medicine : A Journal of the British Diabetic Association 2010;27:1017–23.
[38] Rodriguez-Poncelas A, Garre-Olmo J, Franch-Nadal J, Diez-Espino J, Mundet-Tuduri X, Barrot-De la Puente J, et al. Prevalence of chronic kidney disease in patients with type 2 diabetes in Spain: PERCEDIME2 study. BMC Nephrol 2013;14: 46.
[39] Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020;20:280.
[40] Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007-2017. Cardiovasc Diabetol 2018;17:83.
[41] Vijay K, Neuen BL, Lerma EV. Heart failure in patients with diabetes and chronic kidney disease: challenges and opportunities. Cardiorenal Medicine 2022;12:1–10.
[42] Rossing P, Groop PH, Singh R, Lawatscheck R, Tuttle KR. Prevalence of Chronic Kidney Disease in Type 1 Diabetes Among Adults in the U.S. Diabetes Care 2024; 47:1395–9.
[43] Calvo-Hueros JI, Martin-Hidalgo-Barquero MV, Morales-Gabardino JA, Buitrago F. Chronic kidney disease prevalence and cardiovascular risk in a cohort of patients with type 2 diabetes followed for 10 years in Badajoz (Spain). An Observational Study Primary Care Diabetes 2021;15:391–6.
[44] Zelnick LR, Weiss NS, Kestenbaum BR, Robinson-Cohen C, Heagerty PJ, Tuttle K, et al. Diabetes and CKD in the United States population, 2009-2014. Clin J Am Soc Nephrol 2017;12:1984–90.
[45] Cosentino F, Grant PJ, Aboyans V, Bailey CJ, Ceriello A, Delgado V, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. Eur Heart J 2019;2020(41):255–323.
[46] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
[47] Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: Assessing the impact of missing data in electronic health records. J Biomed Inform 2023;139:104269.
[48] Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol 2021;21: 234.
[49] Herman WH. Diabetes Epidemiology: Guiding Clinical and Public Health Practice: The Kelly West Award Lecture, 2006. Diabetes Care 2007;30:1912–9.
[50] Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, Lindsetmo RO, Kouskoumvekaki I, Girolami M, et al. Sci Rep 2017;7:46226.