Abstract. This paper presents new data about diabetes prevalence and illness duration from a population of patients with diabetes Type 1 (43,818) and Type 2 (457,247). Unlike the usual approach that employs adjusted estimates in similar prevalence reports, this study extracts data from a large number of original clinical documents such as all the outpatient records (6,887,876) issued in Bulgaria to all the 501,065 patients with diabetics during 2018 (9.77% of all the 5,128,172 patients recorded in 2018, 4.43% male and 5.35% female). Diabetes prevalence data are described in terms of distributions of Type 1 and Type 2 diabetes over age and gender. It is mapped to an Observational Medical Outcomes Partnership Common Data Model that is publicly available. The distribution of Type 2 diabetics agrees with peak values of BMI established in related research. A major novelty in this research are the data about the diabetes illness duration. It is an essential metric for evaluating the quality of processes evolving over time. Accurate estimates are obtained for the duration measured in years of Type 1 (95% CI 10.92, 11.08) and Type 2 (95% CI 7.97, 8.02) diabetics from the Bulgarian population. The patients with diabetes Type 1 have longer diabetes duration relatively to those with Type 2. It is recommended to include this metric in official diabetes prevalence reports.

Keywords. Diabetes, prevalence, illness duration, outpatient record, population

1. Introduction

Diabetes prevalence reports are regularly published with the main objective to access and update existing public health prevention strategies by focusing them more precisely on the population groups affected by this illness [1] [2]. Since most countries experience difficulties in sharing and exchanging clinical data among national healthcare providers, most of these reports are prepared by statistical processing of data extracted from questionnaires [3]. The samples of data from questionnaires substitute the limited availability of data sources of original clinical documents. A major challenge in this approach is to satisfy the requirement for randomness in sampling diabetes data.
Otherwise, the statistical procedures might lead to wrong generalizations and inferences about the whole population [4]. The adoption of a standardized Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) allows to compensate the lack of observational data in questionnaires used for diabetes surveillance reports [5][6]. This way, observational data like illness duration, blood sugar, blood pressure levels, body-mass-index enable the application of standardized analytics for the purposes of obtaining population-level estimates as well as patient-level prediction research [7][8]. Quite often observational data is represented in clinical documents in terms of free native language and requires complicated natural language processing of large amounts of text.

The limited data on diabetes prevalence extracted from original clinical documents and in particular, diabetes illness duration, at population level has motivated us to present this research work. The objective is to provide a population-based summary of diabetes prevalence in Bulgaria by the end of 2018 from the viewpoint of the diabetes duration distributed by age and gender categories. Unlike other prevalence reports, data is extracted from original clinical documents such as the outpatient records prepared by general practitioners or medical specialists on every patient’s visit for medical examination or consultation. The dataset of outpatient records is the latest publicly available in Bulgaria. This dataset is nationally representative because it comprises all the outpatient records (6,887,876) issued in Bulgaria to patients with diabetics during 2018. The records are provided by the National Health Insurance Fund (NHIF [9]) in Bulgaria on the request of the Bulgarian Ministry of education and the Ministry of health for the purposes of National scientific program eHealth [10]. The methods and the results from computer experiments are provided in the following section. In Section 3 we summarize the findings in this paper and compare the results with existing data sources.

2. Methods and Results

The outpatient records collected in 2018 have been provided in XML format, pseudonymized and validated against a predefined XML schema. The XML schema was mapped to the relational database model used by the Bulgarian Diabetes Register (BDR [6]), that maintains data about diabetes prevalence in Bulgaria since 2013 [11]. This way we were able to identify 501,065 distinct patients (9.77% of all the patients in 2018, 4.43% male and 5.35% female) with diabetes in 2018 compared to 492,939 in 2017. It is above the average for the USA (9.1%) [1] and above the average (6.2%) for the EU [2].

![Figure 1. Distribution by age of patients with diabetes Type 1.](image)

Figures 1 and 2 display the distribution of diabetics with Type 1 (mean=68) and Type 2 (mean=65). Notice, that the numbers for both male and female patients with Type 2 start...
to increase sharply at age 40-45. For comparison, in a related research [5] we establish at the same age a peak in the number of diabetics in obesity category 1 (BMI>30).

![Figure 2. Distribution by age of patients with diabetes Type 2.](image)

Figures 3 and 4 display the diabetes duration extracted from the medical history using natural language processing. After applying the Interquartile range rule to remove the outliers [4], we get overall skewness, correspondingly 0.4 for Type 1 and 0.5 for Type 2.

![Figure 3. Diabetes illness duration by gender of patients with diabetes Type 2.](image)

![Figure 4. Diabetes illness duration by gender of patients with diabetes Type 1.](image)
Thus, we get quite accurate estimates for the diabetes duration for Type 1 (95% CI 10.92, 11.08) and Type 2 (95% CI 7.97, 8.02) measured in years.

3. Discussion and Conclusions

This paper presents new data about diabetes prevalence and illness duration from a population of patients with diabetes Type 1 (43,818) and Type 2 (457,247). Unlike similar prevalence reports this study extracts data from a large number of original clinical documents such as all the outpatient records (6,887,876) issued in Bulgaria to 501,065 patients with diabetics during 2018 (9.77% of all the 5,128,172 patients recorded in 2018, 4.43% male and 5.35% female). The distributions of Type 1 and Type 2 diabetes over age and gender are calculated entirely from these data records rather from statistical estimates as the usual practice is. The male patients prevail over the female in numbers until age 70, when the total of female patients starts to sharply overwhelm the male. The distribution of Type 2 diabetics agrees with peak values of BMI established in related research. A major novelty in this research are the data about the diabetes illness duration. It is an essential metric for evaluating the quality of processes evolving over time. In particular, we get accurate estimates for the duration (in years) of Type 1 (95% CI 10.92, 11.08) and Type 2 (95% CI 7.97, 8.02) diabetics from the Bulgarian population. The patients with diabetes Type 1 have longer diabetes duration relatively to those with Type 2.

Acknowledgement

This research is supported by grant 80-10-8/11.04.2023 and NSP eHealth in Bulgaria.

References