data

MDPI

Article

Fundamentals of Analysis of Health Data for Non-Physicians

Carlos Hernández-Nava [1,*], Miguel-Félix Mata-Rivera [1] and Sergio Flores-Hernández [2]

[1] Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas del, Gustavo A Madero, Ciudad de México 07340, Mexico; mmatar@ipn.mx
[2] Center for Health Systems Research, National Institute of Public Health, Av. Universidad 655, Col. Santa María Ahuacatitlán, Cuernavaca 62100, Morelos, Mexico; sergio.flores@insp.mx
* Correspondence: hernandeznc@ipn.mx

Abstract: The increasing prevalence of diabetes worldwide, including in Mexico, presents significant challenges to healthcare systems. This has a notable impact on hospital admissions, as diabetes is considered an ambulatory care-sensitive condition, meaning that hospitalizations could be avoided. This is just one example of many challenges faced in the medical and public health fields. Traditional healthcare methods have been effective in managing diabetes and preventing complications. However, they often encounter limitations when it comes to analyzing large amounts of health data to effectively identify and address diseases. This paper aims to bridge this gap by outlining a comprehensive methodology for non-physicians, particularly data scientists, working in healthcare. As a case study, this paper utilizes hospital diabetes discharge records from 2010 to 2023, totaling 36,665,793 records from medical units under the Ministry of Health of Mexico. We aim to highlight the importance for data scientists to understand the problem and its implications. By doing so, insights can be generated to inform policy decisions and reduce the burden of avoidable hospitalizations. The approach primarily relies on stratification and standardization to uncover rates based on sex and age groups. This study provides a foundation for data scientists to approach health data in a new way.

Keywords: methodology; data science; public health; standardization; epidemiology; datasets; hospital discharges; adjusted rate

Citation: Hernández-Nava, C.; Mata-Rivera, M.-F.; Flores-Hernández, S. Fundamentals of Analysis of Health Data for Non-Physicians. *Data* **2024**, *9*, 112. https://doi.org/10.3390/data9100112

Academic Editor: Rüdiger Pryss

Received: 25 May 2024
Revised: 22 September 2024
Accepted: 23 September 2024
Published: 27 September 2024

1. Introduction

Non-communicable diseases account for 74% of all annual deaths worldwide, as reported by the World Health Organization (WHO) [1]. Diabetes mellitus (DM) is classified as an ambulatory care-sensitive condition (ACSC), meaning that effective and timely treatment can help avoid hospitalization. These hospitalizations are closely tied to the quality of ambulatory care, also known as primary care, which is why they are considered potentially avoidable.

Data from INEGI indicate that Mexico ranks among the top countries in the world for the prevalence of diabetes, with approximately 10.3 million adults diagnosed with the condition. Additionally, the burden of avoidable hospitalizations related to complications of diabetes mellitus (AHRDMs) remains a critical issue, resulting in significant economic costs and placing strain on healthcare resources [2]. According to the 10th IDF Diabetes Atlas released by the International Diabetes Federation in 2021, Mexico had 14,123,200 cases of diabetes in adults, ranking eighth in the world. Projections suggest that by 2030, there will be 16.4 million cases of diabetes in adults. The country also ranked eighth in diabetes-related health expenditure [3].

The field of public health focuses on preventing health issues and improving overall population health. There is a common misconception among non-physicians that public health is primarily concerned with making predictions. Another key concept is confounding

variables, which refers to a variable or factor that influences an outcome when a researcher is attempting to measure another factor. This project is intended to establish a groundwork for data scientists who are addressing health-related concerns, ensuring that their insights and conclusions hold significant value.

Traditional healthcare methods have played a crucial role in managing diabetes and averting complications [4,5], yet they are sometimes constrained in effectively identifying and addressing AHRDMs. Data science techniques have recently provided new opportunities to comprehend disease patterns and enhance healthcare delivery [6,7]. Harnessing advanced analytics, artificial intelligence, and big data science represents a promising strategy to support traditional healthcare approaches and enhance results.

Machine learning (ML) and deep learning (DL) have shown great potential in improving outcomes. However, it is crucial to understand how to properly apply biostatistics and regression methods, which serve as the foundation for ML and DL. According to Olawade et al. in [8], traditional methods face challenges in handling diverse and complex data sources, emphasizing the critical role of data scientists in managing these complexities due to their expertise in handling large relational or document databases. This publication serves as a valuable reference for data scientists venturing into this field.

Despite the significant progress in this field, there remains a gap in the current literature concerning incorporating data science techniques to address AHRDMs in patients, particularly within the Mexican healthcare system. Our research is a foundation for integrating data science techniques into healthcare practices. Through this interdisciplinary analysis, we strive to provide valuable insights that can guide policy decisions and, in our specific context, mitigate the impact of AHRDMs in Mexico.

The National Institute of Statistics and Geography of Mexico, in its press release commemorating World Diabetes Day [9], has provided the following statistics: 13% of deaths were attributed to diabetes, with 51% affecting men (71,330) and 49% affecting women (69,396). The mortality rate associated with diabetes was 11 per 10,000 inhabitants. Additionally, 10.3% of the national population aged 20 and above were diagnosed with diabetes. In 2021, the state of Mexico ranked fourth, and Mexico City ranked ninth in terms of mortality rate due to diabetes.

The Agency for Healthcare Research and Quality (AHRQ) has established a health indicator known as the Prevention Quality Indicator (PQI) number 93 [10]. This composite indicator is a rate that is based on hospital discharges with specific principal diagnosis codes from the ICD-10-CM. These codes include cases such as diabetes with short-term complications (PQI #1), diabetes with long-term complications (PQI #3), uncontrolled diabetes (PQI #14), or lower-extremity amputation among patients with diabetes (PQI #16).

Records in all datasets that correspond to any of the PQI #93 specification codes and have the principal diagnosis of avoidable hospitalizations related to diabetes mellitus (AHRDMs) are discussed in the upcoming sections. The denominator refers to the study population, which, in this context, consists of individuals aged 20 and above from Mexico City and the State of Mexico, collectively referred to as the Metropolitan Area (MA) in this article. This geographical area is one of the most densely populated regions globally. As of 2023, the study population stood at 19,219,280.

In Mexico City and the State of Mexico, the AHRDM rates decreased from 48.48 in 2010 to 39.83 in 2019. In 2019, this decrease led to eight women receiving effective and timely treatment, preventing hospitalization due to diabetes mellitus compared to 2010. For men, the number was only 2. This trend continued from January 2020, before the COVID-19 pandemic, to December 2023. Over this period, the AHRDM rates for women were 21.57 in 2020 and 30.55 in 2023, while for men, the rates were 29.56 in 2020 and 28.67 in 2023 (all rates per 100,000 people). Consequently, nine women were unable to avoid hospitalization during the four-year period from January 2020 to December 2023, while the rate for men remained relatively unchanged.

The assessment of AHRDM rates is essential as it provides crucial prevention quality indicators, empowering healthcare professionals to make well-informed decisions about

healthcare systems. This work entails a cross-sectional study design that involves retrospective analysis of secondary databases. The term "retrospective" is used because the study utilizes past records, and "secondary" because the databases were not originally intended for this purpose [11,12]. The records pertain to hospital discharges for adults aged 20 and above with diabetes mellitus between 2010 and 2023.

In addition to this introduction and the abstract, the paper encompasses a section detailing the methodology employed. This section specifies each step, including: data collection, referencing databases from INEGI and SS/DGIS; outlier analysis for the elimination or imputation of data; and computation of age-adjusted rates of AHRDM, direct standardization, and the calculation of specific rates. The results section presents a table detailing the step-by-step process for the rates or the PQI #93 indicator. The discussion section delves into insights, such as highlighting that the highest adjusted rate was for the age group between 45 and 64 years old in the metropolitan area of Mexico City. In theAppendix A, there are details of the cleaning and pre-processing, outlining specific characteristics of the sources and variables selected for the study; and the population of metropolitan area of Mexico City.

2. Methods

The following section provides an overview of the health methodology, data sources, outlier analysis, specific and crude rates, stratification, and direct standardization. Crude rates represent unadjusted rates and are calculated in a straightforward manner by data scientists. Conversely, adjusted-age rates by sex are commonly used in the field of health. Both processes are detailed in this section.

2.1. Methodology

Data science methodology involves five stages, including data collection, data preprocessing, modeling, interpretation of the results, and results communication [13]. Its aim is to derive insights that support informed decision-making through techniques such as outlier analysis, data imputation, and data regression modeling algorithms. However, for public health issues aimed at creating social and health impact, it is essential to adhere to the guidelines established by physicians, particularly epidemiologists. A visual representation of how a data scientist can collaborate on a health science project is shown in Figure 1.

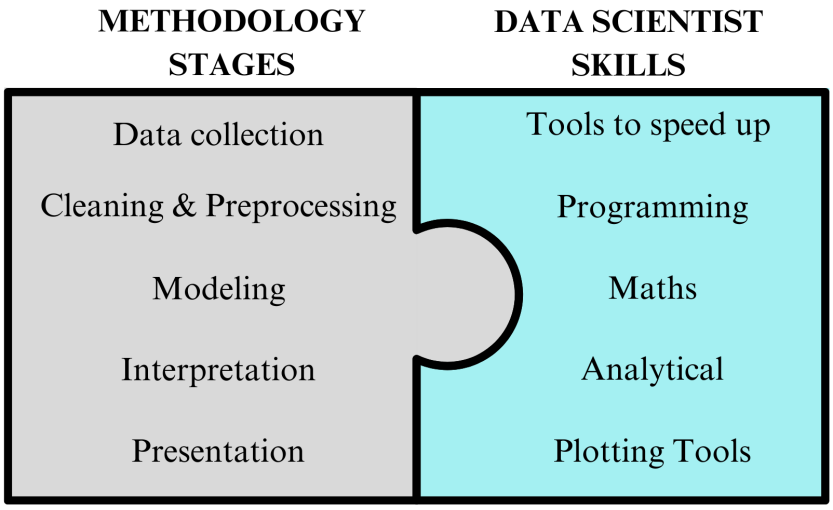

Figure 1. Convergence between health researchers and data scientists: methodology in health-related fields shares similarities with data science. A significant advantage of collaboration between health researchers and data scientists is the efficient and rapid identification of solutions to problems.

Remember to account for confounding variables, especially in public health. One way to address them during data analysis is through stratification [14]. In our study, we stratify by two potential confounding variables: age and sex. Data standardization [15,16] is crucial for ensuring that results can be compared and conclusions drawn across different locations, countries, and study populations. Standardized values enable other scientists to use the data without being influenced by geographical or population characteristics [17,18].

In the realm of public health methodology, there are five crucial stages. Data collection is reliant on Electronic Health Records (EHR) as the primary data source, supplemented by surveys and interviews. Data scientists must navigate large and diverse datasets during this stage, often dealing with data of questionable quality. Cleaning and pre-processing are essential due to the significance of standardization in public health. This involves the stratification and adjustment of data directly or indirectly. Managing missing values, data imputation, and analyzing outliers are fundamental tasks requiring strong programming skills.

Modeling primarily involves biostatistical methods and regression algorithms in public health. As Goldstein et al. refer to in [19], data scientists with strong analytical thinking skills can significantly enhance this process. Interpretation in public health projects requires a focus on biological and clinical implications rather than generic data science metrics. Mistakes in this stage can occur due to a lack of understanding of public health foundations. Presentation is reliant on statistical evidence, such as confidence intervals from statistical tests or biostatistical methods. Data scientists proficient in visualization tools play a crucial role in this stage, and collaboration with a physician is essential for effective presentation.

2.2. Data Sources

In this study, the primary data source utilized is the open dataset provided by The Ministry of Health of Mexico through the General Department of Health Information (SS/DGIS). This dataset includes open datasets of deaths, hospital discharges, births, maternal deaths, and emergencies. The analysis focuses on hospital discharge records from 2010 to 2023, encompassing a total of 36,665,793 records from medical units under the Ministry of Health of Mexico. Table 1 displays the national records by year, the post-cleaning process records, the corresponding records of the MA, filtered AHRDM cases, and the population of 20 years and older. Also, it presents the percentages in each column in brackets relative to the universal value reported in the previous column.

Table 1. Hospital discharge records for Mexico and the Metropolitan Area of Mexico City (MA) were obtained from open datasets provided by SS/DGIS. These records include corresponding population data sourced from an open dataset from INEGI, and are broken down per year. The dataset includes both values and percentages where applicable.

Year	National Records	After Cleaning	MA Records	MA Diabetes Cases	Population Habs. *
2010	2,634,339	2,632,251 (99.92%)	510,888 (19.41%)	7071 (1.38%)	15,160,396
2011	2,775,189	2,774,330 (99.97%)	536,111 (19.32%)	8089 (1.51%)	15,472,618
2012	2,880,706	2,880,075 (99.98%)	566,765 (19.68%)	7455 (1.32%)	15,784,839
2013	2,879,313	2,879,052 (99.99%)	576,107 (20.01%)	7109 (1.23%)	16,097,061
2014	2,959,197	2,958,924 (99.99%)	603,358 (20.39%)	7074 (1.17%)	16,409,284
2015	2,970,812	2,970,483 (99.99%)	581,673 (19.58%)	6675 (1.15%)	16,721,507
2016	2,955,144	2,952,697 (99.92%)	599,528 (20.30%)	7478 (1.25%)	17,033,727
2017	2,729,341	2,715,873 (99.51%)	535,983 (19.74%)	7538 (1.41%)	17,345,950
2018	2,623,379	2,622,560 (99.97%)	530,078 (20.21%)	7530 (1.42%)	17,658,172
2019	2,629,434	2,628,771 (99.97%)	515,396 (19.61%)	8329 (1.62%)	17,970,393
2020	1,937,344	1,934,458 (99.85%)	387,942 (20.05%)	4767 (1.23%)	18,282,615
2021	2,088,780	2,088,352 (99.98%)	391,540 (18.75%)	4371 (1.12%)	18,594,837
2022	2,203,636	2,197,685 (99.73%)	372,418 (16.95%)	5643 (1.52%)	18,907,058
2023	2,399,179	2,390,869 (99.65%)	381,771 (15.97%)	6011 (1.57%)	19,219,280
Total	**36,665,793**	**36,626,380** (99.89%)	**7,089,558** (19.36%)	**95,140** (1.34%)	**240,657,737**

* People of 20 years and older.

2.3. Outlier Analysis

In order to obtain the AHRDM rates, the initial step involves cleaning outliers, imputing data if necessary, or deleting records. Subsequently, only the cases of MA were selected. When conducting outlier analysis, it is crucial to utilize hypothesis testing to identify any unusual data points. In this particular analysis, records containing patients aged "999" and those with sex codes outside the allowed coding (1 for men and 2 for women) were removed. Such analysis is vital, as databases often harbor missing, redundant, or spurious values that have the potential to yield inaccurate results. The "After Cleaning" column in Table 1 indicates that 99.89% of the records are valid.

2.4. Age-Adjusted Rates

The avoidable hospitalization related diabetes mellitus (AHRDM) rate is calculated by dividing the total cases by the study population of 20 years and older, resulting in what is known as the crude rate. The process starts with collecting discharge hospitalization records from SS/DGIS, followed by a cleaning and filtering process to match cases between MA discharge hospitalizations and the PQI #93 ICD-10-MD code. These counted cases serve as the numerator, and when divided by the MA population of 20 years and older, provide the crude rate (per 100,000 population).

Even crude rate is useful because it provides the real measure of a problem. When comparing with other geographical areas, regions, or countries, it is essential to recognize that crude rates can be deceptive as they do not account for specific population structures and characteristics, such as sex and age. However, the use of crude rates depends on the specific rate being calculated; for instance, crude mortality rate is commonly used. Moving from top to bottom, the process of calculating the crude rate is illustrated in Figure 2. Subsequently, the data are divided by sex and age groups to calculate age-specific rates for women and men. Age-adjusted rates are then derived using the standard population as weights. This involves utilizing a standard population with a typical age structure, such as the country's population. The age groups considered are 20–44 years (young adults), 45–64 years (adults), and 65 years and older (elderly adults), further divided into men and women.

The primary purpose of standardization is to compare rates across time and geographical areas. For instance, in the European Union, the proportion of people 65 years and older increased from 16% to 21% from 2002 to 2022 [20]. This demographic trend is relevant when comparing rates with the study population's proportion of people 65 years and older, which is 10%, half of the European Union's proportion.

Sex-specific crude rates of AHRDM are computed by the specific study population, dividing the cases by men and women. For example, in 2020, there were 2173 cases of AHRDM in women and 2594 cases in men; if these values are divided by the specific study population of women and men, the specific rate for women is 22.55 AHRDMs, and for men it is 30 AHRDMs, both per 100,000 habs (see Table 2 for each year).

In computing rates, the data were stratified by sex, but the population also has biological characteristics by age group. Therefore, to customize the analysis, records are divided by age group: 20–44, 45–64, and 65 years and older. Analyzing rates by age and sex enables us to compare across years, accounting for potential changes in the population structure, although these differences are likely minimal. Table A1 shows the diabetes cases and populations per age group.

In order to standardize the data and calculate specific rates, it is necessary to adjust for age groups and sex (see Table 3). The process involves multiplying the crude rate by the proportion (weight) of the standard population to get the specific rate. This step ensures that the data are properly adjusted or weighted. The adjusted rates, along with a 95% confidence interval, are presented in the fourth column, labeled "age & sex-adjusted". The lower and upper interval limits are shown in parentheses, and the error is indicated in brackets. These values were computed in STATA V18 [21] using the command `dstdize`.

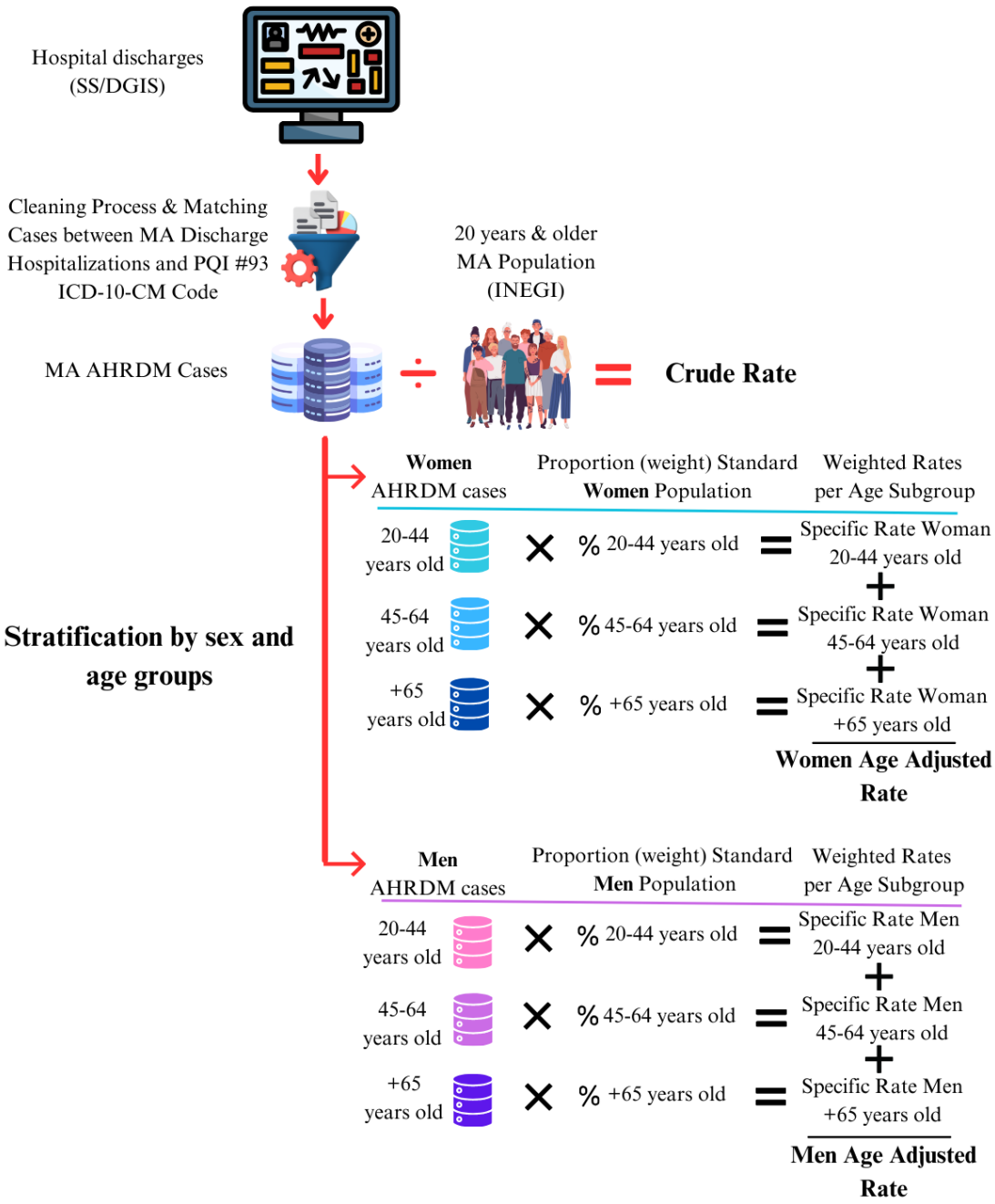

Figure 2. Crude and age-adjusted rates step-by-step calculation: first, merge the dataset of hospital discharges from SS/DGIS. Then, clean the data and select diabetes-related cases. Add the population data from INEGI to obtain the crude rate. To get age-adjusted rates, stratify by sex and divide each stratum into three subgroups by age. Multiply the cases by a standard population proportion (weight) to obtain a specific rate per age group. Finally, calculate the age-adjusted rates by adding the specific rates for women and men, accordingly.

Table 2. Crude rates over the years per 100,000 inhabitants were obtained by dividing the number of diabetes-related hospitalizations (cases) by the study population of women and men in Mexico City and the State of Mexico.

Year	Cases		Population (Habs.)		Crude Rates (×100,000 Habs.)	
	Women (A)	Men (B)	Women (C)	Men (D)	Women (A/C)	Men (B/D)
2010	3611	3460	8,010,757	7,149,639	45.08	48.39
2011	4001	4088	8,173,368	7,299,250	48.95	56.01
2012	3360	3795	8,335,979	7,448,860	43.91	50.95
2013	3544	3565	8,498,589	7,598,472	41.70	46.92
2014	3354	3720	8,661,201	7,748,083	38.72	48.01
2015	3276	3399	8,823,812	7,897,695	37.13	43.04
2016	3565	3913	8,986,422	8,047,305	39.67	48.63
2017	3544	3994	9,149,034	8,196,916	38.74	48.73
2018	3678	3852	9,311,644	8,346,528	39.50	46.15
2019	3943	4386	9,474,255	8,496,138	41.62	51.62
2020	2173	2594	9,636,866	8,645,749	22.55	30.00
2021	2023	2348	9,799,477	8,795,360	20.64	26.70
2022	2607	3036	9,962,088	8,944,970	26.17	33.94
2023	3314	2697	10,124,698	9,094,582	32.73	29.65

Table 3. Age-adjusted rates of AHRDM with a confidence level of 95% and their corresponding lower and upper bounds. Crude and specific rates per year and age-group are used to calculate the final age-adjusted rate.

Year	Crude Rate		Specific Rate *		Age and Sex Adjusted CI 95% (Lower,Upper) [Error]	
Years-Old	Women	Men	Women	Men	Women	Men
2010						
20–44	15.49	19.76	8.98	11.46	48.48	53.52
45–64	80.47	93.85	24.01	31.03	(46.90,50.07)	(51.71,55.32)
+65	127.97	115.83	14.03	15.45	[0.81]	[0.92]
2011						
20–44	17.44	22.46	10.12	13.03	51.58	60.86
45–64	88.86	109.03	26.66	32.71	(49.98,53.18)	(59.02,62.78)
+65	123.36	126.03	14.8	15.12	[0.82]	[0.96]
2012						
20–44	16.65	19.69	9.66	11.42	45.47	54.71
45–64	77.56	98.51	23.27	29.55	(43.99,46.94)	(52.97,56.47)
+65	104.47	114.51	12.54	13.74	[0.75]	[0.89]
2013						
20–44	15.60	19.02	9.05	11.03	42.61	49.47
45–64	70.68	90.40	21.20	27.12	(41.21,44.02)	(47.84,51.10)
+65	103.03	94.36	12.36	11.32	[0.72]	[0.83]
2014						
20–44	13.61	18.16	7.89	10.53	39.06	50.45
45–64	65.39	86.89	19.62	26.07	(37.73,40.38)	(48.82,52.07)
+65	96.23	115.40	11.55	13.85	[0.67]	[0.83]
2015						
20–44	14.11	16.47	8.18	9.55	36.98	44.51
45–64	58.86	79.32	17.66	23.80	(35.71,38.25)	(43.01,46.01)
+65	92.81	93.02	11.14	11.16	[0.65]	[0.77]
2016						
20–44	16.51	19.29	9.58	11.19	39.07	49.65
45–64	63.28	88.56	18.98	26.57	(37.79,40.35)	(48.09,51.21)
+65	87.57	99.12	10.51	11.89	[0.65]	[0.79]

Table 3. *Cont.*

Year	Crude Rate		Specific Rate *		Age and Sex Adjusted CI 95% (Lower,Upper) [Error]	
Years-Old	Women	Men	Women	Men	Women	Men
2017						
20–44	14.62	18.56	8.48	10.76	37.69	49.21
45–64	62.90	89.86	18.87	26.96	(36.45,38.93)	(47.68,50.74)
+65	86.18	95.76	10.34	11.49	[0.63]	[0.78]
2018						
20–44	16.28	17.79	9.44	10.32	38.12	46.14
45–64	62.81	84.20	18.84	25.26	(36.89,39.35)	(44.68,47.60)
+65	81.96	88.02	9.84	10.56	[0.63]	[0.74]
2019						
20–44	17.52	21.61	10.16	12.53	39.83	51.15
45–64	64.76	91.32	19.43	27.40	(38.58,41.07))	(49.63,52.66)
+65	85.34	93.48	10.24	11.22	[0.64]	[0.77]
2020						
20–44	11.44	14.85	6.64	8.61	21.57	29.56
45–64	33.31	49.13	9.99	14.74	(20.66,22.48)	(28.43,30.70)
+65	41.19	51.77	4.94	6.21	[0.46]	[0.58]
2021						
20–44	10.02	10.88	5.81	6.31	19.56	25.98
45–64	29.93	48.01	8.98	14.40	(18.71,20.42)	(24.93,27.03)
+65	39.74	43.89	4.77	5.27	[0.44]	[0.54]
2022						
20–44	11.69	14.98	6.78	8.69	24.53	32.91
45–64	38.53	56.84	11.56	17.05	(23.57,25.46)	(31.74,34.08)
+65	51.48	59.73	6.19	7.17	[0.48]	[0.60]
2023						
20–44	14.99	13.10	8.70	7.60	30.55	28.67
45–64	52.52	44.45	15.76	13.33	(29.51,31.60)	(27.58,29.75)
+65	50.85	64.43	6.10	7.73	[0.53]	[0.55]

* Specific rate is computed by multiplying crude rate by the proportion of the standard population: 20–44 is 0.58, 45–64 is 0.3, +65 is 0.12.

3. Results

By collaboration between a data scientist and a health researcher, adjusted rates were obtained quickly with the use of programming scripts and specialized software. These rates were validated by an epidemiologist. Figure 3 shows the age-adjusted rates (red for men and blue for women).

The age-adjusted rate of avoidable hospitalizations for diabetes is higher in men than in women from 2010 to 2022. However, in 2023, the rate is higher for women. Regardless of age, there was a 50% decrease in sex-adjusted rates from 2019 to 2020. Additionally, the age-adjusted rate was higher than the crude rate for women in 2010, but from 2011 to 2023, the crude rate was higher. After cleaning and outlier analysis, 36,626,380 (99.89%) records remain, with 0.11% outliers deleted over the years.

Before the pandemic period, from 2010 to 2019, adjusted rates for men were higher. In Figure 3, the pink line represents the crude rate, illustrating the impact of COVID-19 from 2020 to the present.

Figure 4 presents a visual comparison of specific crude rates and specific rates. The solid lines denote male data, while the dashed lines represent female data. In the 20–44 age group, both male and female populations exhibit lower adjusted rates. In the 45–64 age group, the adjusted rates also display a decrease, with a more pronounced variance, reaching up to 50 AHRDMs in certain years. For individuals aged 65 and older, the adjusted rates consistently demonstrate a reduction, with some instances showing a variance of over 100 AHRDMs. These findings highlight the criticality of standardization

for specific rates. Figure 5 complements this by delineating the disparities between the crude rates and the specific rates, calculated as the highest rate minus the other rate.

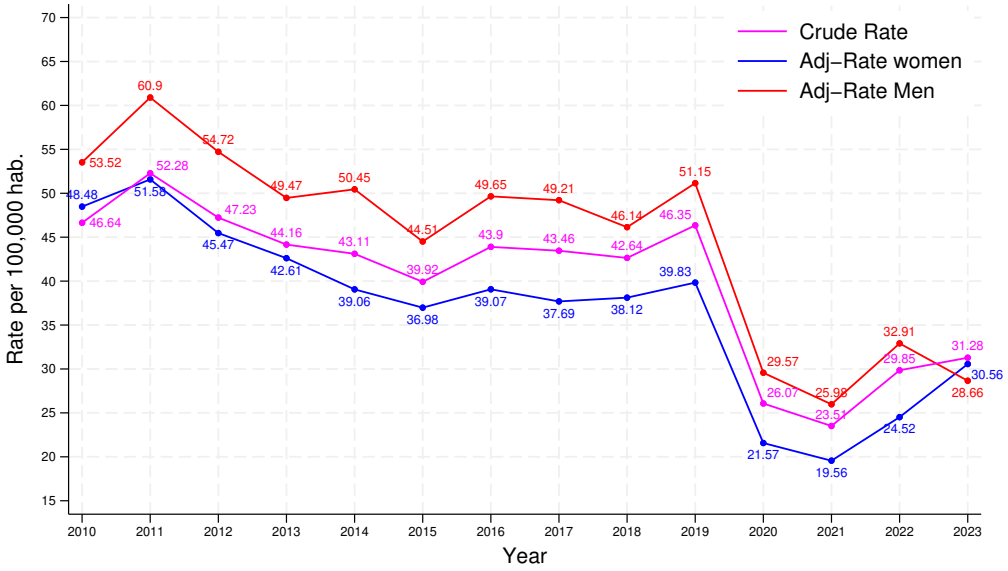

Figure 3. Crude and age-adjusted rates for women and men (per 100,000 habs.). Cases of AHRDM and population of the Metropolitan Area were stratified by three age groups: 20–44, 45–64, and 65 years and older.

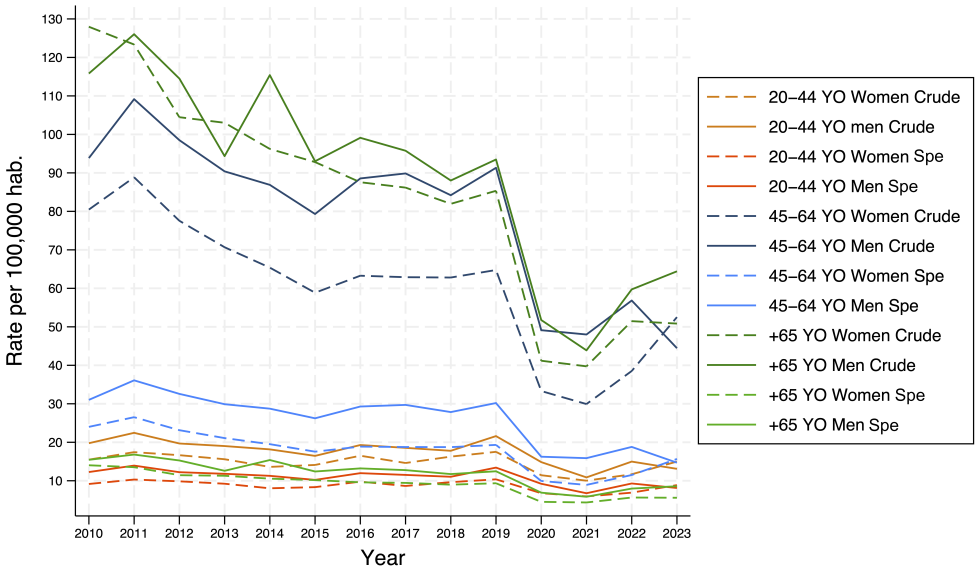

Figure 4. The comparison between crude and specific rates per age group over the years illustrates the impact of stratifying by age group for both women and men. It is evident that individuals aged 65 and older, as well as those aged 45–64 years old, have contributed to the health problem with crude rates exceeding 100 AHRDMs in certain years, both prior to and during the COVID-19 pandemic. However, upon stratification, specific rates by age group are found to be under 40 AHRDMs.

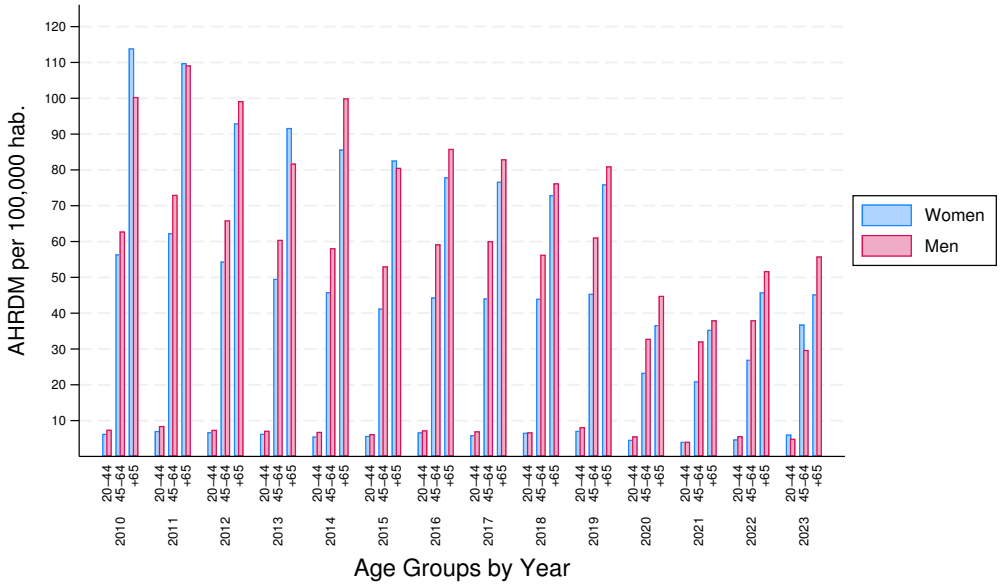

Figure 5. Differences in crude and specific rates per age group by sex over the years. Each blue bar represents the variance in AHRDMs for women between the two rates. Over the years, women aged 65 years and older consistently had over three times the number of AHRDMs compared to those aged 20–44 years old. This trend is also observed for men, indicated by the red bars. However, in the 20–44 age group, variance is almost imperceptible, with less than a 10 point difference in AHRDMs in all years.

Pillars for a Health Data Scientist

A new way to deal with health databases involved five pillars: first, the capacity to manage diverse data sources and formats; second, knowledge of the variables and the foundations of the field where data science is applied; third, the foundations of biostatistics and script programming; fourth, superb skills to manage software and specialized languages to clean, impute, analyze, and everything related to process data; and finally, the interpretation of the results, which will only be valuable when a data scientist is involved with the field of application.

4. Discussion

In our discussions about the involvement of a non-physician in initiatives to address health issues, under the guidance of an epidemiologist who provided the conceptual framework of public health, we successfully calculated rates. Crude rates offer a clear understanding of the true extent of the AHRDM issue, and we also derived adjusted rates of AHRDM for comparison with other regions or countries.

Both rates serve as a benchmark for gauging the efficiency of diabetes prevention programs in Mexico City and State of Mexico. This procedure can be reproduced by altering the origin of hospitalization data and categorizing the population based on demographic characteristics, typically age groups and sex.

In this case study, if neither stratification nor standardization rates were applied, it would not be possible to compare with other research works. It is imperative for the data scientist to consider the necessity of stratification and standardization, as the applicability of findings may vary across different demographic and geographical contexts. Epidemiology serves as a valuable branch of public health for examining various social determinants, including but not limited to poverty and access to health services, and for evaluating the functionality and efficiency of healthcare systems in order to enhance the effectiveness of current prevention policies for diabetes-related complications.

5. Conclusions

It is essential for non-physicians without a medical background to acquaint themselves with the methods, processes, techniques, and developments within the healthcare field in order to effectively address the issue at hand. Even if a non-physician, such as a data scientist, feels adequately prepared to interpret the results, it is crucial to bear in mind that research is constantly progressing, and it is vital to collaborate with experts from various fields to ensure accurate implementation and interpretation of results.

The similarity between the crude rate and the age-adjusted rate for women indicates that the rate of hospitalization is increasing uniformly across all age groups. This suggests a widespread trend of rising hospitalization rates for women, regardless of their age. In the year 2023, there has been a noticeable rise in the rate of AHRDMs in women. This calls for attention to conduct comprehensive studies on the quality of care provided to women affected by this condition, warranting a thorough investigation into the root causes and potential solutions. It has become evident that there may be deficiencies in the prevention of diabetes-related complications, specifically in the female population.

Due to the fragmented nature of the Mexican health system, the population with social security and other private services was not taken into account. Indeed, 35% of Mexicans are affiliated to the National Welfare Institute (INSABI, in Spanish) of the Ministry of Health of Mexico, while IMSS (Mexican Institute for Social Insurance), ISSSTE (Institute for Social Security and Services for State Workers), and private health services have 65%. Subsequent research should prioritize the comparison of methodologies for various diseases, including those beyond the realm of healthcare.

Author Contributions: Conceptualization, methodology, supervision, S.F.-H. and M.-F.M.-R.; software, formal analysis, investigation, resources, data curation, writing—original draft preparation, C.H.-N.; validation, S.F.-H.; visualization, writing—review and editing, S.F.-H., M.-F.M.-R. and C.H.-N.; project administration, funding acquisition, M.-F.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.dgis.salud.gob.mx/contenidos/basesdedatos/Datos_Abiertos_gobmx.html# (accessed on 22 September 2024). Also, an example of pre-processed data are available here: https://github.com/carloshn29/datasets (accessed on 22 September 2024).

Acknowledgments: Thanks to CONAHCyT for the Ph.D. studies grant received during this research. Thanks to Secretaría de Investigación y Posgrado del Instituto Politécnico Nacional and postgraduate and research section of the Interdisciplinary Professional Unit in Engineering And Advanced Technologies, UPIITA-IPN.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ICD-10-CM	International Classification for Diseases, 10th revision, Clinical Modification
AHRDMs	Avoidable Hospitalizations Related to Diabetes Mellitus
DGIS	Dirección General de Informacion de Salud (due to terms of use, acronym must be in Spanish
AHRQ	Agency for Healthcare Research and Quality
ACSC	Ambulatory Care-Sensitive Condition

PQI	Prevention Quality Indicator
MA	Metropolitan Area of Mexico City
SS	Secretaría de Salud (due to terms of use, acronym must be in spanish)

Appendix A. Population of the Metropolitan Area of Mexico City

Data Cleansing and Pre-Processing

To make the study reproducible, it should be mentioned that the language used to clean and pre-process the data is R, with R Studio as the IDE to load data depending on the year of the file; even if all datasets are CSV files, each file has its characteristics, which is the reason why the delimiter symbol and the number of variables (columns of the tables) differ from each other. For example, the separator symbol is the comma for the 2010 file year, giving the following code sentence:

```
read.csv(file = ‘‘data/open/EGRESO_2010.csv’’, header = TRUE, sep = ’,’)
```

In the same example, the number of variables is: `11,13,17,18,39` which correspond to the variables `EDAD`, `SEXO`, `ENTIDAD`, `MUNIC`, and `AFECPRIN`, consecutively. `EDAD` is the patient age, `SEXO` is the patient sex, 1 for men and 2 for women, `ENTIDAD` is the state name of the patient, `MUNIC` is the municipality name where the patient resides, and `AFCPRIN` contains the ICD-10-CM registered at hospital discharge.

The cleaning process includes a filtering process using regular expressions of the correct ICD-10-CM codes. We take only the first four alphanumeric values with the regular expression in the sentence `grepl(‘‘[A-Z][0-9][0-9][A-Z0-9]’’, AFECPRIN)`.

Table A1 shows in its columns, first, the year and age groups, second, the number of MA cases and percentages by men and women, and third, the MA populations of men and women for each age group.

Table A1. Diabetes cases of the Metropolitan Area of Mexico City over the years and corresponding populations by age groups.

Year	Metropolitan Area Cases		Population (Habs.)	
Years-Old	Women	Men	Women	Men
2010	3611 (100%)	3460 (100%)	8,010,757	7,149,639
20–44	768 (21.3%)	903 (26.1%)	4,958,925	4,569,823
45–64	1800 (49.8%)	1841 (53.2%)	2,236,828	1,961,667
+65	1043 (28.9%)	716 (20.7%)	815,004	618,149
2011	4001 (100%)	4088 (100%)	8,173,368	7,299,250
20–44	869 (21.7%)	1034 (25.3%)	4,983,308	4,602,965
45–64	2069 (51.7%)	2227 (54.5%)	2,328,376	2,040,086
+65	1063 (26.6%)	827 (20.2%)	861,684	656,199
2012	3360 (100%)	3795 (100%)	8,335,979	7,448,860
20–44	834 (22.8%)	913 (24.1%)	5,007,691	4,636,107
45–64	1877 (51.3%)	2087 (55.0%)	2,419,923	2,118,504
+65	949 (25.9%)	795 (20.9%)	908,365	694,249
2013	3544 (100%)	3565 (100%)	8,498,589	7,598,472
20–44	785 (22.2%)	888 (24.9%)	5,032,073	4,669,249
45–64	1775 (50.1%)	1986 (55.7%)	2,511,471	2,196,923
+65	984 (27.8%)	691 (19.4%)	955,045	732,300
2014	3354 (100%)	3720 (100%)	8,661,201	7,748,083
20–44	688 (20.5%)	854 (23.0%)	5,056,456	4,702,391
45–64	1702 (50.7%)	1977 (53.1%)	2,603,019	2,275,342
+65	964 (28.7%)	889 (23.9%)	1,001,726	770,350
2015	3276 (100%)	3399 (100%)	8,823,812	7,897,695
20–44	717 (21.9%)	780 (22.9%)	5,080,839	4,735,534
45–64	1586 (48.4%)	1867 (54.9%)	2,694,567	2,353,761
+65	973 (29.7%)	752 (22.1%)	1,048,406	808,400

Table A1. *Cont.*

Year Years-Old	Metropolitan Area Cases Women	 Men	Population (Habs.) Women	 Men
2016	3565 (100%)	3913 (100%)	8,986,422	8,047,305
20–44	843 (23.6%)	920 (23.5%)	5,105,222	4,768,676
45–64	1763 (49.5%)	2154 (55.0%)	2,786,114	2,432,179
+65	959 (26.9%)	839 (21.4%)	1,095,086	846,450
2017	3544 (100%)	3994 (100%)	9,149,034	8,196,916
20–44	750 (21.2%)	891 (22.3%)	5,129,605	4,801,818
45–64	1810 (51.1%)	2256 (56.5%)	2,877,662	2,510,598
+65	984 (27.8%)	847 (21.2%)	1,141,767	884,500
2018	3678 (100%)	3852 (100%)	9,311,644	8,346,528
20–44	839 (22.8%)	860 (22.3%)	5,153,987	4,834,960
45–64	1865 (50.7%)	2180 (56.6%)	2,969,210	2,589,017
+65	974 (26.5%)	812 (21.1%)	1,188,447	922,551
2019	3943 (100%)	4386 (100%)	9,474,255	8,496,138
20–44	907 (23.0%)	1052 (24.0%)	5,178,370	4,868,102
45–64	1982 (50.3%)	2436 (55.5%)	3,060,757	2,667,435
+65	1054 (26.7%)	898 (20.5%)	1,235,128	960,601
2020	2173 (100%)	2594 (100%)	9,636,866	8,645,749
20–44	595 (27.4%)	728 (28.1%)	5,202,753	4,901,244
45–64	1050 (48.3%)	1349 (52.0%)	3,152,305	2,745,854
+65	528 (24.3%)	517 (19.9%)	1,281,808	998,651
2021	2023 (100%)	2348 (100%)	9,799,477	8,795,360
20–44	524 (25.9%)	537 (22.9%)	5,227,136	4,934,386
45–64	971 (48.0%)	1356 (57.8%)	3,243,853	2,824,273
+65	528 (26.1%)	455 (19.4%)	1,328,488	1,036,701
2022	2607 (100%)	3036 (100%)	9,962,088	8,944,970
20–44	614 (23.6%)	744 (24.5%)	5,251,519	4,967,528
45–64	1285 (49.3%)	1650 (54.3%)	3,335,400	2,902,691
+65	708 (27.2%)	642 (21.1%)	1,375,169	1,074,751
2023	3314 (100%)	2697 (100%)	10,124,698	9,094,582
20–44	791 (23.9%)	655 (24.3%)	5,275,901	5,000,670
45–64	1800 (54.3%)	1325 (49.1%)	3,426,948	2,981,110
+65	723 (21.8%)	717 (26.6%)	1,421,849	1,112,802

References

1. Noncommunicable Diseases. Available online: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (accessed on 18 April 2024).
2. Salas-Zapata, L.; Palacio-Mejía, L.S.; Aracena-Genao, B.; Hernández-Ávila, J.E.; Nieto-Lopez, E.S. Costos Directos de las hospitalizaciones por diabetes mellitus en el Instituto Mexicano del Seguro Social. *Gac. Sanit.* **2018**, *32*, 209–215. ISSN 0213-9111. [CrossRef] [PubMed]
3. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed.; International Diabetes Federation: Brussels, Belgium, 2021; ISBN 978-2-930229-98-0.
4. Agudelo, M.; Murillo, J.; Gutierrez, L.; Giraldo, L. *Hospitalizaciones y Muertes Evitables por Condiciones Sensibles a Atención Primaria en Salud. México, 2005–2014*; Consejo Nacional de Población: Mexico City, Mexico, 2017.
5. Flores, S.; Acosta, O.; Hernández, M.I.; Delgado, S.; Reyes, H. Calidad de la atención en diabetes tipo 2, avances y retos de 2012 a 2018–2019 para el sistema de salud de México. *Salud Publica Mex.* **2020**, *62*, 618–626. [CrossRef] [PubMed]
6. Wood, S.M.; Yue, M.; Kotsis, S.V.; Seyferth, A.V.; Wang, L.; Chung, K.C. Preventable Hospitalization Trends before and after the Affordable Care Act. *AJPM Focus* **2022**, *1*, 100027. [CrossRef] [PubMed]
7. Saxena, A.; Ramamoorthy, V.; Rubens, M.; McGranaghan, P.; Veledar, E.; Nasir, K. Trends in quality of primary care in the United States, 2007–2016. *Sci. Rep.* **2022**, *12*, 1982. [CrossRef] [PubMed]
8. Olawade, D.; Wada, O.; Kunonga, E.; Abaire, O.; Ling, J. Using artificial intelligence to improve public health: A narrative review. *Front. Public Health* **2023**, *11*, 1196397. [CrossRef] [PubMed]
9. INEGI (Mexico), Press Release No. 657/22, 10 November 2022 .
10. Agency for Healthcare Research and Quality. Prevention Quality Indicator 93 (PQI 93), Prevention Quality Diabetes Composite. In *AHRQ Quality Indicators ICD-10-CM/PCS Specification*; Agency for Healthcare Research and Quality: Rockville, MD, USA, 2023.

11. Rattanavipapong, W.; Wang, Y.; Butchon, R.; Kittiratchakool, N.; Thammatacharee, J.; Teerawattananon, Y.; Isaranuwatchai, W. Retrospective secondary data analysis to identify high-cost users in inpatient department of hospitals in Thailand, a middle-income country with universal healthcare coverage. *BMJ Open* **2021**, *11*, e047330. [CrossRef] [PubMed]
12. Goode, W.; Punjabi, V.; Niewiara, J.; Roberts, L.; Bruce, J.; Silva, S.; Morgan, B.; Pereira, K.; Brysiewicz, P.; Clarke, D. Using a Retrospective Secondary Data Analysis to Identify Risk Factors for Pulmonary Complications in Trauma Patients in Pietermaritzburg, South Africa. *J. Surg. Res.* **2021**, *262*, 47–56. ISSN 0022-4804. [CrossRef] [PubMed]
13. Cao, L. Data Science: A Comprehensive Overview. *ACM Comput. Surv.* **2017**, *50*, 43. [CrossRef]
14. Jager, K.J.; Zocalli, C.; MacLeod, A.; Dekker, F.W. Confounding: What it is and how to deal with it. *Kidney Int.* **2008**, *73*, 256–260. [CrossRef] [PubMed]
15. Naing, N.N. Easy way to learn standardization: Direct and indirect methods. *Malays. J. Med. Sci.* **2000**, *7*, 10. [PubMed] [PubMed Central]
16. Higham, J.; Flowers, J.; Hall, P. Standardization. *Inf. Public Health Obs. Recomm. Methods* **2005**, *6* , 1–8. ISSN: 1477-7290.
17. Merchant, A.T. Standardization. In *Statistical Approaches for Epidemiology*, 3rd ed.; Mitra, A.K., Ed.; Springer: Cham, Switzerland, 2024; pp. 147–154. [CrossRef]
18. Keiding, N.; Clayton, D. Standardization and Control for Confounding in observational studies: A Historical Perspective. *Stat. Sci.* **2014**, *29*, 529–558. [CrossRef]
19. Goldstein, N.D.; LeVasseur, M.; McClure, L.A. On the Convergence of Epidemiology, Biostatistics, and Data Science. *Harv. Data Sci. Rev.* **2020**, *2*. [CrossRef]
20. Noncommunicable Diseases. Available online: https://ec.europa.eu/eurostat/web/interactive-publications/demography-2023 (accessed on 18 April 2024).
21. StataCorp. *Stata Statistical Software: Release 18*; StataCorp LLC: College Station, TX, USA, 2023.