Recovery from the decrease in the incidence of chronic diseases in primary care during the COVID pandemic: analysis of time series in the primary care practices of the Catalan Institute of Health

Roser Cantenys Sabà

Master's Thesis
Degree in Data Science and Engineering
Université de Bordeaux
Institut Català de la Salut

June, 2022

Address:
Université de Bordeaux
Institut de Santé Publique d'Epidémiologie et de Développement
146 rue Léo Saignat
CS 61292
33076 Bordeaux cedex
www.u-bordeaux.fr

Acknowledgements

First of all, I would like to thank Manuel Medina for giving me the opportunity of doing my master's thesis in Institut Català de la Salut, and for all the help that he has given me during these months.

Second, I would want to express my gratitude to Ermengol, my thesis advisor, for helping me throughout the process and teaching me how to do epidemiological research and analysis from both a theoretical and practical standpoint.

I would want to thank my co-director, Núria, in particular, for working with me on some of the topics and issues of my thesis, as well as for all of her patience and hours spent introducing me to the most methodological parts. I would also like to thank Carol to the hours she spent reading and commenting on my work. And thank you to each one of the members of SISAP's team for taking an interest in my work, sharing their advise and opinions and making me feel like I was a part of the group from the very first day.

Finally, I would want to thank my family for all of their support and assistance throughout the years, especially my parents for instilling in me a love of learning, knowing, and discovering.

Abstract

Objectives

Examine whether there is a recovery from the decline in the incidence of chronic disease diagnoses during COVID-19's second year, March 2021-March 2022, to see if primary care has regained diagnostic detection at levels similar to or higher than before the pandemic, or if additional measures are needed to ensure the diagnosis of certain chronic diseases.

Methods and materials

A retrospective observational research was undertaken on 5 million patients aged 14 and above. The diagnoses were taken from the Catalan primary care electronic health records. The average monthly incidence from 2014 until the pandemic outbreak was compared with the period of COVID-19 (March 2020 - March 2022) for the following chronic diseases: malignant neoplasms, type 2 diabetes mellitus, ischemic heart disease, and chronic obstructive pulmonary disease. Analysis: Time series analysis was utilised for descriptive analysis, time series regression was used to compare expected and observed incidences, and the Welch test was used to compare difference of means. The difference was considered significant if the p-value was less than 0.05. Ethical aspects: approved by the corresponding ethics committee.

Results

All chronic illnesses were diagnosed at a lower rate in 2020: -31% for malignant neoplasms, -48% for diabetes, and -55% for chronic obstructive pulmonary disease. However, during 2021, the incidence of some pathologies recovered to levels similar to those before the pandemic, such as malignant neoplasms. For others, an increase in diagnoses was observed, such as diabetes up 34% in the second year of COVID-19 compared to pre-pandemic years or ischaemic heart disease. In contrast, diagnosis of chronic obstructive pulmonary disease continued at lower levels than in the pre-pandemic years (-43%).

Conclusions

Throughout the pandemic, and especially in 2020, the number of chronic illness diagnoses has plummeted drastically. The occurrence of these diagnoses has varied over the course of COVID-19's second year. It has rebounded to levels close to or greater than before the pandemic for a variety of illnesses, illustrating primary care's adaptability; nonetheless, several diseases remain below pre-pandemic levels.

Keywords— COVID-19, Chronic diseases, Neoplasms, Ischemic heart disease, Diabetes, Chronic Obstructive Pulmonary disease, Chronic diseases incidence.

Contents

1 Introduction

1.1 Background

Coronavirus disease 2019, also known as COVID-19, is a disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) that first emerged as an outbreak in Wuhan, China in December 2019 and swiftly expand into a global pandemic [1]. On January 24, 2020, the first cases in Europe were confirmed in France [2], and the first recognized case in Catalonia was on February 25, 2020.

COVID-19 has been declared a global public health emergency by the World Health Organisation (WHO) on January 30, 2020, and has wreaked havoc on global health systems, with ramifications in all facets of human life that we know [3, 4] due its high transmissibility [5]. SARS-CoV-2 has had a significant influence on public health, in addition to the virus's positive cases, as it has disrupted the usual functioning of health care services. It has resulted in a considerable reduction in non-COVID hospital admissions [6, 7], non-critical non-COVID-19 elective surgeries [8], and emergency room visits, resulting in changes in daily health practice [9]. At the beginning of the pandemic, in the absence of a viable vaccine, the rapid spread and the novelty of the disease, nonpharmaceutical interventions (NPI) such as hand hygiene, social distancing, and, in certain cases, house confinement were utilised to reduce transmission. Most countries established measures of incarceration and social distancing [10, 11] in the initial wave, and Spain expressly adopted a national lockdown on March 14, 2020 [12].

Those NPIs have impacted on access to screening tests and on the long-term care of patients with chronic conditions globally as many people postponed their appointments due to government stay-at-home orders, hospital recommendations pushing patients to postpone non-urgent visits, changes in health-seeking behaviour of patients due to the fear of being infected by SARS-CoV-2. In June 2020, 41% of adults in the United States had postponed or avoided medical treatment; 12% emergency care and 32% routine appointments [13]. Other studies such as one conducted at St. Paul's hospital millennium medical college in Ethiopia showed that, overall, the essential services were affected by the COVID-19 pandemic and specifically the inpatient admission showed a 73.3% reduction from the pre-COVID period [14]. Other countries such as Germany also recorded a decrease of physician consultations [15].

Globally, an online survey posted 31 March to 23 April 2020 targeted at healthcare professionals, where 47 countries responded, aimed to evaluate the global impact of COVID-19 on routine care for chronic diseases. Virtual communication was reported as the most common modification in normal treatment. Diabetes, chronic obstructive pulmonary disease, and hypertension were the conditions most affected by the lack of access to care. During COVID-19 pandemic, 80% of respondents said their patients' mental health deteriorated. To avoid an increase in non-COVID-19-related morbidity and mortality, the authors stand that it is critical that normal care persist despite the pandemic [16].

In a study conducted in Catalonia's central area [17], it was observed that new diagnoses decreased by 31% on average in 2020 compared to 2019, with significant reduc-

tions in April (61%), May (56%), and November (52%) with a higher rate of decrease in neoplasms (50%).

All the avoided or postponed visits caused by the COVID-19 have led to a significant impact on the number of cases recorded as well as on the disminution of tests in a vast sort of pathologies all around the world, such as highlighted in the literature for HIV in 44 countries in 4 continents [18] or Borreliosis in Poland [19]. All these facts could have negative consequences for people who did not receive or delay the suitable treatment.

Non-communicable diseases (NCD) kill 41 million people each year, accounting for 71% of all deaths worldwide [20]. They have been recognized for long as the leading cause of mortality and disability worldwide [21]. NCD preventive and treatment services have been significantly affected since the COVID-19 pandemic began, according to a WHO study done in 155 countries [22], since health services have been partially or fully suspended in several countries. More than half of the countries surveyed, have reduced or eliminated hypertension treatment; 49% have reduced or eliminated diabetes and diabetes-related complications treatment; 42% have reduced or eliminated cancer treatment; and 31% have reduced or eliminated cardiovascular emergency treatment.

Physical separation or isolation can lead to poor management of NCD behavioural risk factors such as poor diet, inactivity, cigarette use, and hazardous alcohol use. Chronic illnesses might worsen without effective management owing to stressful situations arising from restrictions, unstable economic situations, and changes in usual health behaviours, according to evidence from this and past pandemics. Physical separation, restricted access to primary health care units, pharmacies, and community services, as well as a reduction in transportation linkages, all impair continuity of care for NCD patients, as they do with other health services and preventative programs. In NCD patients, the disruption of routine health care and medical supplies jointly with the disruption in patient screening, treatment and surveillance of NCD increases the risk of morbidity, impairment, and unnecessary mortality over time [21]. Furthermore, it affects those with pre-existing NCDs differently and may result in de novo NCD sequelae. This pandemic will almost certainly have long-term consequences that will influence practitioners and patients in this sector for years to come [23].

In particular, cancer diagnoses have dropped dramatically in a number of countries [24, 25, 26]. In Belgium, diagnosis of malignant neoplasms reduced by 44% in April 2020 compared to the same month in 2019 [27], while in Catalonia, diagnosis of malignant neoplasms decreased by 34% from March to September 2020 compared to the expected according to previous years [28].

Various studies published in 2021 have demonstrated how the breakout of COVID-19 and related control efforts can wreak havoc on the most vulnerable people, worsening the state of patients with noncommunicable diseases and jeopardising the health-care system's long-term viability [29]. Several professionals have pointed out the need to carry out psychological and screening approaches as well as to develop new methods for the monitoring of the disease and adapt therapeutic strategies in the post-COVID era to not miss patients with a chronic disease and new cases who were undiagnosed during the

COVID pandemic [30, 31, 32]. Meanwhile, we must address the basic needs of NCD patients by defining priority activities to assist them in managing their chronic diseases, both during the current emergency and in the medium/long term. To protect vulnerable individuals and reduce the outbreak's impact, we need to develop an integrated plan that includes both private firms and institutional bodies. Politicians at all levels, including national and European, must make this a priority in their planning plans both during and after the pandemic [33].

1.2 Institution

All this project is carried out in the Institut Català de la Salut (ICS, Catalan Institute of Health) in Sistema d'Informació dels Serveis d'Atenció Primària (SISAP, Primary Care Information System) department.

The ICS, as a reference entity in our country's public health system, works to enhance people's health and quality of life through the promotion of healthy behaviours, the prevention of health issues, and the treatment of diseases, ranging from the most minor to the most complicated.

With 41,000 professionals servicing nearly six million people across Catalonia, the ICS is the largest public health services company in the country. Hence, the main primary care provider in Catalonia. The ICS oversees 283 primary care practices. In other words, it manages around 75% of all primary care practices in the Catalan public health system and covers around 5.8M people from the 7.5M of the catalan inhabitants. Its population is highly representative of the population of Catalonia in terms of geographical area and age and gender distributions [34].

In 2005, ICS implemented a EHR system for use in primary care, known as ECAP (Estació Clínica d'Atenció Primària), a software system that serves as a repository for structured data on diagnoses (coded according to the International Classification of Diseases 10th revision, ICD-10, and the ICD-10-CM adapted version made by WHO), clinical variables, prescription data, laboratory test results and diagnostic requests. These data from ICS' primary care EHR have been previously validated in several studies [**34**].

In addition to healthcare, the ICS conducts extensive research through its seven research institutions, which are integrated within its hospitals and primary care facilities. The study conducted by ICS experts tries to answer problems emerging from clinical practice, with the goal of using the scientific findings to improve people's health. The same may be said about teaching. In Catalonia, the ICS is a pioneer in the education of tomorrow's professionals. Its centres educate 2,300 health-care professionals in 51 specialisations and 4,500 undergraduate students in medicine, nursing, dentistry, and other fields. In addition, the Institute creates continuing education programs for many types of professionals. Overall, to maintain ICS's position as a premier organisation serving residents.

SISAP, leadered by the physician Manuel Medina, took its first steps in 2006, with the aim of providing information, mainly for clinical management, to different professionals and management structures. Since then, SISAP has been obtaining and processing data

to construct useful indicators to a variety of health-care professions and management groups as well as storing data to give support to research projects.

This project may be conducted thanks to all of the tasks provided by SISAP and its resources. From this we expect to provide data on the potential recovery from the decrease in the incidence of diagnoses due to various chronic diseases during the COVID-19 pandemic as well as a comparison of the pre-pandemic and current incidence of these diseases.

1.3 Objectives and hypothesis

Although there is a lot of literature about the decrease of new diagnoses, as we have stated in the background section, we have not found any study that analyses whether the number of diagnoses after almost two years of pandemic is similar before it. Therefore, the goal of this project is to see if there is a recovery from the decline in the incidence of chronic disease diagnoses during the second year of COVID-19, March 2021 - March 2022, in order to see if primary care has regained diagnostic detection at levels similar to or higher than before the pandemic, or if some measures are required to ensure the diagnosis of certain chronic diseases.

It has been shown that the rebound from the lowering frequency of chronic diseases in primary care is linked with the COVID-19 pandemic [3]. We hypothesise that after living with the pandemic for more than two years now, the system could have recovered the diagnosing capacity to levels similar to the ones before the pandemic. Perhaps, this capacity is even higher than before the COVID-19 outbreak since the diagnoses that were missed during the pandemic are being detected now.

Furthermore, we would like to investigate the characteristics of the recovery of chronic illness incidence, such as the patient's socioeconomic status, age, sex, or rurality. We hypothesise changes in the number of diseases diagnosed based on patient variables.

On the other hand, we would also like to question the difference between the recovery of the incidence rate of each pathology. We hypothesise that diagnosis of diseases with simpler diagnostic tests, such as diabetes (which only needs a blood test to be detected) are going to be more recovered than diagnoses of pathologies with more complicated diagnostic tests, such as some types of neoplasm, which require image and blood tests. However, maybe we could observe that when there are facilities and policies in place to diagnose some disease such as breast cancer on a regular basis, the number of instances discovered may recover faster. Therefore, maybe diagnostic tests and policies limit the health system's diagnostic detection.

If this study confirms the recovery of the decline in the incidence of diagnoses related to chronic diseases, then it could be argued that primary care did a good job adapting to the new circumstances diagnosing chronic diseases while not neglecting non-COVID-19 visits. If this diagnostic recovery is not possible due to a chronic illness, these findings can be extremely useful in planning the necessary measures to strengthen primary care in order to ensure the continuation of their care work and, in this case, to improve the screening, diagnosis, and management of certain chronic diseases.

2 Methods

2.1 Study design

We perform a longitudinal retrospective study of chronic case reported diagnosis in ICS primary care health practices. We carry out a retrospective population based study where we start from individual data and then we aggregate them by age, diagnoses, healthcare area, socioeconomic status, sex and year and month of diagnoses to obtain the correspondent aggregated data.

2.2 Study period

The study period of this project was from 1st January 2014 and 28th February 2022. In order to perform the study, this period has been divided into 4 subperiods:

- Training set: 1st January 2014 - 31st December 2018
- Validation set: 1st January 2019 - 28th February 2020
- Analysis sets:
  - **First year of pandemic:** 1st March 2020 - 28th February 2021. Including the state of alarm in Spain which meant lockdown and different phases of de-escalation and social distancing as well as the very first post lockdown period.
  - **Second year of pandemic:** 1st March 2021 - 28th February 2022. This period is characterised by COVID-19 vaccination that started on 27th December 2020 in Catalonia jointly with softer restrictions regarding social distancing.

2.3 Participants

This study includes all the patients assigned to the primary care practices of the ICS older than 14 years old at the time of diagnosis with at least one of the following non-communicable disease diagnoses registered in the EHR: Malignant Neoplasms, Diabetes, Chronic Obstructive Pulmonary Disease, Ischemic Heart Disease registered in the EHR and made by primary care professionals from January 2014 until December 2021. All the ICD-10 codes used for these diagnoses can be found in Appendix A.

2.4 Variables

2.4.1 Monthly incidence rate

The main variable that we want to study in this project is the monthly incidence rate of the following non-communicable diseases: Malignant Neoplasms, Diabetes, Ischemic Heart Disease and Chronic Obstructive Pulmonary Disease. Note that the monthly incidence rate is calculated per 100,000 inhabitants as follows:

$$\text{Monthly incidence rate} = \\ = \frac{\text{number of new diagnoses in a month}}{\text{number of alive people at the last day of the month assigned to ICS}} \times 100,000 \quad (1)$$

It is important to note that these diagnoses are just made by primary care professionals, therefore, those diseases detected in hospitals or private centres are not going to be evaluated.

2.4.2 Outcomes

From our main variable, three different outcomes will be also studied: diagnoses' recovery, percentage of diagnoses drop and percentage of excess of diagnoses. If the incidence rate for any study period is similar to that of the training period, we will consider that diagnoses have recovered. On a similar line, we consider the percentage of excess and drop of diagnoses as the deviation percentage of the observed diagnoses during the COVID-19 years in comparison with the historical data. The mathematical definition of these outcomes will be explained later in section Statistical methods.

2.4.3 Stratification variables

We believe that there exist variables that may significantly impact the percentage diagnosis' recovery, one of our outcome variables. Hence, we have divided the target population into stratums. The stratification variables taken into account to create these subgroups are disease, sex, age at diagnoses date, socioeconomic status and rurality.

Concerning diseases, malignan neoplasms, diabetes, ischemic heart disease and chronic obstructive pulmonary disease are considered. As previously stated, ICD-10 codes have been used to define the diagnoses and they can be found in Appendix A. It is important to note that different ICD-10 codes can codify for the same disease. Therefore, we have a set of codes that codify each of the studied diseases, not just one.

Regarding the age, different age groups have been created categorising the variable into 4 groups defined by quartiles. All these groups will be calculated for each of the studied diseases. This way, each stratum will have the same amount of data. Hence, we will obtain four homogeneous groups.

On the other hand, concerning socioeconomic status, we used a validated deprivation indicator (MEDEA deprivation index) [35] based on census data to determine socioeconomic status. This index is constructed by the project Mortality and socio-economic and environmental inequalities in small areas of cities in Spain [35]. Instead of utilising the MEDEA Deprivation Index as a continuous variable, we opted to divide it into quartiles, with the first and fourth quartiles representing the least and most disadvantaged districts, respectively.

Finally, rural areas were categorised separately and were defined as areas with less than 10,000 inhabitants and a population density lower than 150 inhabitants/km^2.

In a nutshell, the information we analyse will be aggregated by age (four categories), sex (two categories), disease (four categories), and socioeconomic status plus rurality (five categories).

2.5 Data sources

All data was obtained from the primary care EHR of the ICS. Data has been retrieved as aggregated count data on fields: sex, age, health region, month and year of the diagnoses of the previously mentioned diseases. In order to do so, various databases were evaluated. In the next subsections, we will describe the data extraction process and the tables involved for obtaining data related to diagnoses, patients and socioeconomic status.

2.5.1 Diagnoses

First of all, we used diagnoses from a table that contains all of the ICD10-coded health conditions from 1st January 2014 until 28th February 2022 and that includes 238,709,518 diagnoses.

From this table, diagnoses, date of diagnosis (month and year) and centre where it was detected are taken. The considered diagnoses are those codified with ICD10 codification and can be distinguished by the flag *CODE_O_PS* with value = 'C'. There are some centres that have their own code for certain conditions and these are not going to be included.

It is important to note that this table includes all the modifications performed on the diagnosis. For example, if at the beginning a patient was first diagnosed with stomach ache but it ended up being a neoplasm the diagnosis is changed and marked with a flag called *PR_HIST*. This variable will be 1 if the diagnosis is the latest, otherwise it will be a 0. Following with the last example, the stomach ache entry would have *PR_HIST* = 0 and the neoplasm *PR_HIST* = 1. From this database we will just take into account the diagnoses with *PR_HIST* = 1 as we consider that the last modification contains the correct diagnosis.

Then, we also filter the conditions that have not been misdiagnosed, i.e., that have no value at the *TERMINATION_DATE* field.

Finally, all of the diseases we're looking into are chronic. As a result, if a patient has been diagnosed with diabetes twice, or with any other chronic condition, only the first diagnosis will be considered. In the case of neoplasms, if a patient is diagnosed with the same neoplasm twice, there must be a 400-day gap between the diagnosis; otherwise, only the first case will be accepted.

2.5.2 Patients

On the one hand, we have extracted personal information such as date of birth and sex from a table that includes both demographic data and the primary health care center to which the patients were assigned depending on where they lived during the year of the diagnoses. It is important to note that this table includes all the population assigned to a primary health care centre since January 2014. Therefore, it includes 19,574,954 patients (people who have passed away are also recorded into this table).

Thus, sex and date of birth are extracted from this table and used by the analysis.

On the other hand, an historical patients table has been used in order to extract the number of patients assigned to each primary health care centre per month and year. This information has been extracted grouped by sex, age, year, month and primary health care center. Regarding age, as there are few people over 95 years old they all have been considered in the age range of 95 or older.

It is important to note that when the diagnoses table does not provide us the centre where the diagnosis was performed we will use the later historical table in order to impute these values. The aforementioned imputation is performed using the primary health centre where the patient was assigned at the time of diagnosis.

2.5.3 Socioeconomic status

Then, another table which relates the primary health centre where the patients are assigned with the socioeconomic level of the area was employed to represent socioeconomic data.

Therefore, from this table just the MEDEA indicator is used [35]. Rural and urban areas have different behaviours. Studies performed by ICS always use MEDEA levels for urban areas and a single aggregation for rural areas. Therefore, from now on we will consider the socioeconomic variable as a five level variable that includes the four percentiles extracted from the MEDEA indicator for urban areas and the rural characteristic.

2.5.4 Final Database

Note that all this process is shown in Figure 1 and Figure 2 shows the star schema of the data warehouse.

In order to extract, and therefore obtain, the desired information, SQL accessing MariaDB databases and Pyhton 3.10.2 were used to combine all of the databases and count the number of people with the same characteristics, such as age, sex and health areas, who were diagnosed with a certain condition in the same year and month. Because age and health regions where people are assigned might change over time, they were calculated at the time of diagnosis. Also note that each health region corresponds to a specific socioeconomic level, therefore at the end this catalogue is used in order to impute the socioeconomic status. As a result, an aggregated database including aggregated data is obtained at the end of the process.

Although the databases used in this study are really complete and accurate, it is worth noting that that they are not without missing values. Despite the availability of all dates of birth and sexes for all patients assigned to a Catalan primary health centre, the centre to which the patient was diagnosed is not always recorded and sometimes needs to be filled with the centre where he was assigned at the time of diagnosis. Hence, data preprocessing and cleaning have been applied to confirm and check the data quality.

Later, on preprocessing all the data sources are combined in order to calculate incidence rates and obtain the final desidered database that includes the monthly incidence rate by month and year of the diagnoses, disease, age of the people at diagnose, sex, primary health center where the condition was diagnosed, MEDEA of the center and whether

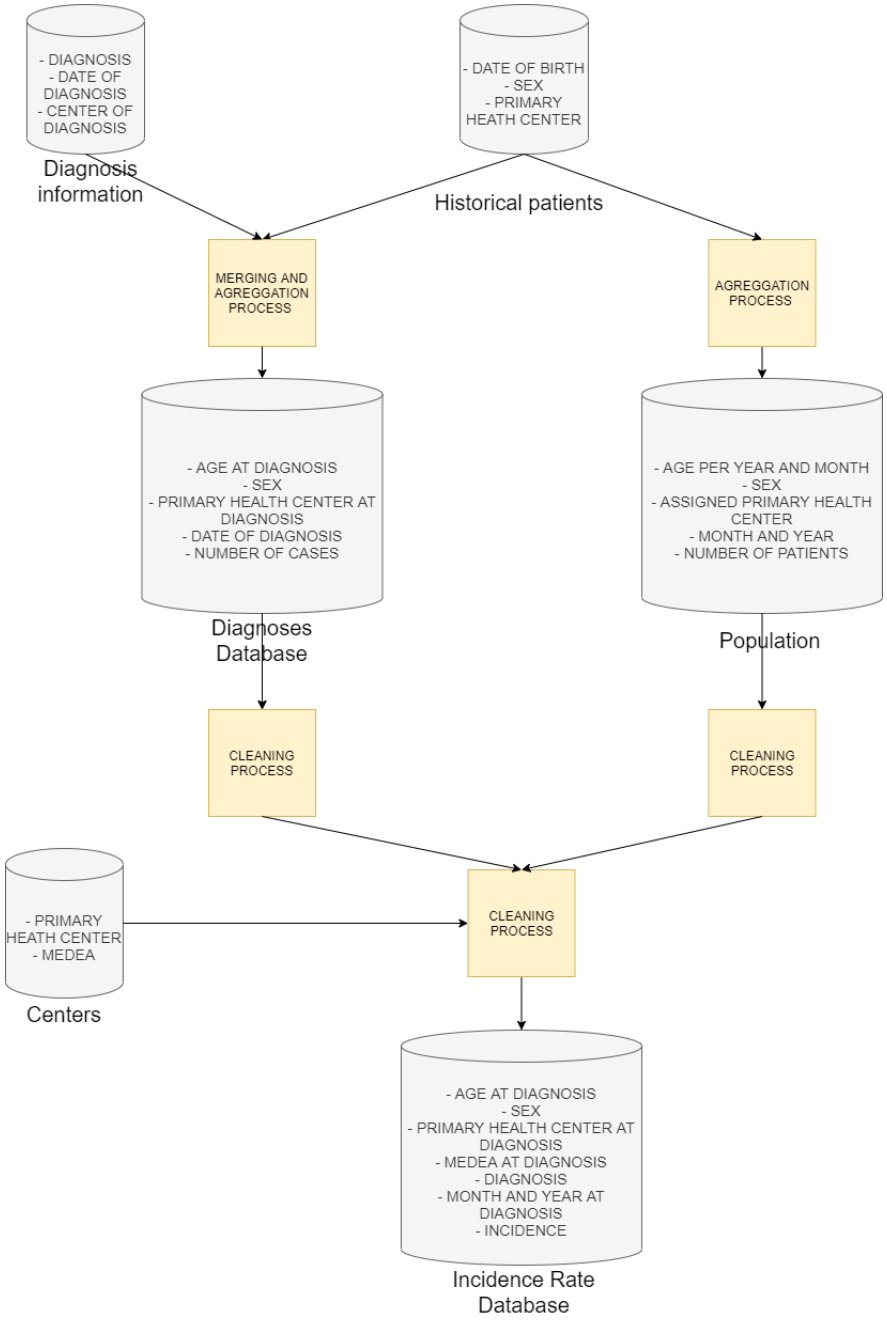


Figure 1: Extract transform and load process.

the center belongs to ICS or not. The final database is saved in a MariaDB server with primary key the combination of date, diagnoses, age, sex and primary health center variables.

The performed preprocessing is explained in Appendix B. This annex includes the flow charts for each of the diseases as well as the preprocessing output obtained per neoplasms. The rest of the diagnoses have been preprocessed and analysed in the same way. However, the outputs are not included there.

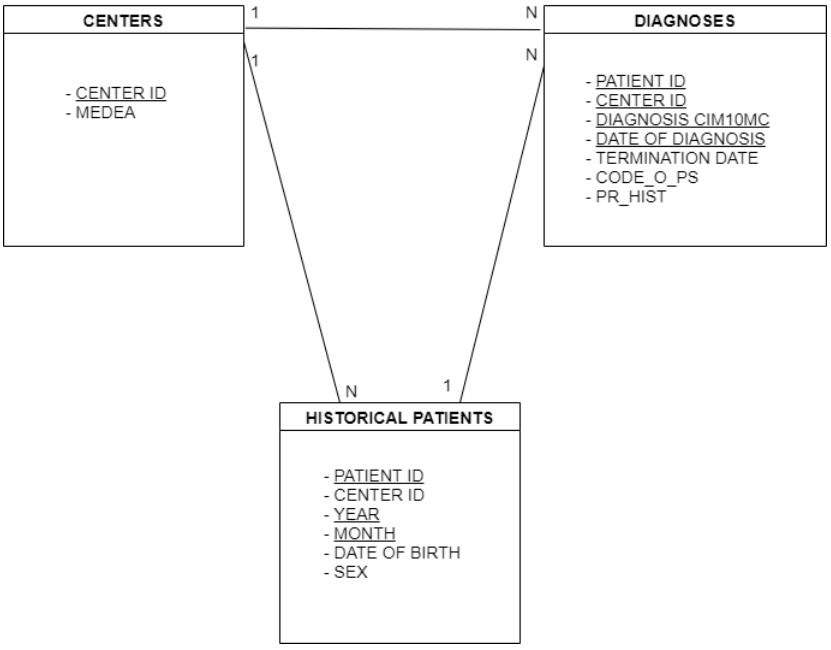


Figure 2: Star schema of the data warehouse.

2.6 Statistical methods

In this project we analyse the monthly incidence rate of four diseases.

Depending on the nature of the data, several methodologies have been used to analyse this variable and acquire the intended conclusions, such as determining whether or not a diagnostic recovery occurs. Time series and Welch test are two of these methodologies.

On the one hand, time series allow us to perform a descriptive analysis over the 8 years compressed in the study in terms of trend, seasonality, and cycles. On the other hand, this method also allows us to perform time-series regression [36] and predict the incidence rate of each disease since March if no pandemic existed. By doing so, we are able to compare the predicted incidence with the observed one.

We obtained the expected incidence and the IC95% for validation and study periods using time series linear regression where the response variable is the incidence rate per 10^5 inhabitants.

As previously stated, the data set was divided into four sections: training (from 2014 to 2018), validation (2019), and analysis of the first year of pandemic March 2020 to March 2021 and second year of pandemic March 2021 to March 2022. As a sensitivity study, we used the training set to adjust the models and the validation set to test our techniques. We looked to see if our methodology detected any increases or decreases in monthly incidence rates in a year that wasn't affected by the COVID-19 pandemic and we validated our models in this period. Finally, once the model was validated we retrained our model including the validation period in the study set and at the end we applied the predicted time series to our datasets for analysis.

For each disease and time period, the number of predicted diagnoses was calculated by multiplying the estimated incidence by the population and dividing by 100,000. The number of observed minus expected diagnoses, computed monthly and exclusively for the periods where observed incidence was below the lower 95 percent CIs of the time series, was used to identify excess or shortage of diagnoses. The diagnoses' recovery is determined as follows:

$$\text{Diagnoses' recovery} = \\ = \frac{\text{number of months the observed incidence rate falls inside the CI95\% of the forecast series}}{\text{length of the period}} \times 100 \tag{2}$$

The percentage of excess or shortage of diagnosis is calculated as the difference between the average observed incidence rate and the expected incidence rate divided by the expected incidence:

$$\text{\% of excess or shortage} = \\ = \frac{\text{observed incidence – expected incidence}}{\text{expected incidence}} \times 100 \tag{3}$$

Before being able to use this strategy, we will need to make sure that the model is validated. In order to validate the model we will check that it fits correctly the validation year, as we have previously mentioned, and in addition that its assumptions are met. We will go deeper on the assumptions in the Time series regression subsection.

In addition, it is important to note that although we have defined the training period as the period between January 2014 and January 2019, before training with all the period we will check the data quality, taking into account outliers, checking the stability of the trend and also seasonalities. If some of the years included in this period are statistically different from the rest or there are some outliers we may consider to exclude some years from the training period or apply some mechanism to minimise the impact of the outliers.

If in any case we cannot validate the time series regression, even cleaning the study period, or making some transformation to data, as for example applying differentiation, we will consider using another descriptive test to define the diagnoses' recovery. In this case, Welch test is used in a descriptive analysis to compare the average monthly incidence rate of the study periods with the baseline, the prepandemic period. In this respect, we are going to conclude that there exists diagnoses' recovery if the comparison between the study period and the one prepandemic are not statistically different. In this scenario we are not going to define the percentage of excess or shortage of diagnosis but in case the means are statistically different, we will perform two extra tests, one for excess and the other one for shortage. For this reason two extra null hypothesis are going to be defined, one stating than the average incidence rate of period A is higher than the one of period B and the second one stating that the average incidence rate of period A is lower than the one of period B.

2.6.1 Time-series Regression

Anything that is observed sequentially over time is a time series. When forecasting time series data, the aim is to estimate how the sequence of observations, in our case, the monthly incidence rate, will continue into the future.

Time series regression is a statistical method for modelling responses based on response history (also known as autoregressive dynamics) and dynamics transfer from relevant predictors. We can use this model to infer unseen data, and in this case to predict what would have happened during the COVID-19 years if the pandemic hadn't happened.

Here we use the trend and the seasonality of the time series as adjustment variables and we will fit a model for each of the stratums we have defined for each diagnose, i.e, one per each sex, one per each age, one per each socioeconomic status and a global one. Therefore, the shape of each model regression will be as follows:

$$Y_t = \beta_0 + \sum_{i=1,..,11} \beta_i \times season_{i,t} + \beta_{12} \times trend + error_t. \tag{4}$$

In order to validate the model there are some assumptions that must be met regarding the residuals (35):

- They must have a mean of zero, else the forecasts will be consistently skewed. For each t, the expected value of the error e_t, given the explanatory variables for all time periods, is zero. Mathematically, $E(e_t|X) = 0$, for t $= 1, 2, ..., n$
- They are uncorrelated; otherwise, there will be additional information in the data that would not have been exploited in the current model. Conditional on X, the errors in two different time periods are uncorrelated: $Corr(e_t, e_s|X) = 0$, for all t! $= s$.
- They have constant variance. Conditional on X, the variance of the errors is the same for all t: $Var(e_t|X) = Var(y_t|X) = \sigma^2$, for t $= 1, 2, ..., n$.
- They are normally distributed. The errors are independent of X and are independently and identically distributed as $Normal(0, \sigma^2)$.

Once the models are validated, we will proceed to forecast the time series in order to compare what actually happened with what we expected. To obtain the predictions the estimated coefficients in the regression equation are used with setting the error term to zero.

2.6.2 Mean Comparison

In order to compare whether the average monthly incidence rate is equal between two periods, the Welch test is used [37]. The Welch test is calculated taking the difference between the sample means, and then dividing it by some estimate of of the standard error of that difference as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE(\overline{X}_1 - \overline{X}_2)} \tag{5}$$

The null hypothesis of this test assumes that both samples have the same mean whereas the alternative hypothesis assumes different means. These hypotheses are not without the assumption that both samples are drawn from a normal population and they are independent. However, the alternative hypothesis does not require the two populations to have equal variance. It is important to note that a p-value of 0.05 is used.

In case we reject the null hypothesis of the Welch test, and in order to be able to conclude if the mean incidence rate for a specific period is higher or lower than the one from another period, we will just compare numerically the two means.

2.7 Ethics

The study is conducted according to the guidelines of the Declaration of Helsinki, last modification, Fortaleza, Brazil 2013 [38] and in accordance with the European Data Protection Regulations and the Spanish Organic Law on Data Protection and Guarantee of Digital Rights. Although analyses are conducted on aggregated data, this work is included in a study which has the ethical approvals from the Clinical Research Ethics Committee of the IDIAP JGol (project code: 20/172-PCV).

3 Results

3.1 Population

In this section we'll go through the characteristics of people assigned to primary care practices, which totaled 7,117,880 people in December 2021. Nonetheless, we will only consider people over the age of 14, with a population of roughly 6M people.

We use the population assigned to Catalan primary care practices in order to calculate the incidence rate for each disease. In table 1 we can see the structure of this population, which has remained stable between the study period and the validation and analysis sets in terms of age, sex and socioeconomic status. Therefore, as can be observed, little over 51% of the population are women. In terms of socioeconomic status around 34% of the population lives in rural areas whereas 66% lives in the urban area. Regarding urban areas, 21% of the population lives in the least deprived zones, 15% in areas belonging to the 2nd quartile of MEDEA indicator, 21% in areas belonging to the 3rd quartile of MEDEA indicator and finally 18% in the most deprived zones.

Table 1: Baseline characteristics of population being studied by training set (2014-2018), validation set (January 2019 - February 2020), first year of pandemic analysis set (March 2020 - February 2021) and second year of pandemic analysis set (March 20201 - February 2022).

	Training set	Validation set	1st analysis set	2nd analysis set
Population older than 14 years	4,833,717	4,912,832	4,934,404	4,984,302
% of women	51.21	51.16	51.13	52.89
Age: % of population younger than 60 years	70.92	70.81	70.65	70.58
Age: % of population between 60 and 70 years	13.76	13.70	13.84	13.96
Age: % of population between 70 and 80 years	8.15	8.56	8.56	8.62
Age: % of population older than 80 years	7.17	6.92	6.95	6.84
Socioeconomic status: % of people in the first quartile (least deprived)	21.70	21.70	21.68	21.70
Socioeconomic status: % of people in the second quartile	15.19	15.25	15.29	15.23
Socioeconomic status: % of people in the third quartile	20.93	20.72	20.72	20.66
Socioeconomic status: % of people in the fourth quartile (most deprived)	18.23	18.46	18.45	18.32
Socioeconomic status: % of people in rural areas (Rural)	23.94	23.87	24.08	24.10

3.2 Diagnoses

In this section we'll present the results of the studied diagnoses: neoplasms, diabetes, ischemic heart disease and chronic obstructive pulmonary disease. First of all, we will take an overview of the monthly incidence rate per disease in the period included in this project, January 2018 - February 2022, to see the global evolution of it. After that, we will expose the results obtained for each disease.

Figure 3 shows the incidence rate per 100,000 inhabitants for all the diseases with respect to the training, validation and study sets. We can see that monthly incidence rates during the training and also validation periods have a more or less constant trend. On the other hand, during the first period of study we can observe a big drop in the incidence rates for all the diseases that are more or less mitigated during the second year of COVID-19. During the later period, we can observe that there is a studied disease that have an incidence rate similar to the one before the pandemic, neoplasms; there are some that obtain a higher incidence rate during the second year of COVID-19 in comparison with what we observed before the pandemic, ischemic heart disease and diabetes; and finally there is one that still has an incidence rate lower than the expected, chronic obstructive pulmonary disease.

We will now go over the characteristics of each disease's incidence in greater depth.

3.2.1 Neoplasms

Before the COVID-19 pandemic in Catalonia, from January 2014 to February 2020, 257,747 new malignant neoplasms were registered in primary care. This represents an average monthly incidence rate of 72.81 cases and 72.98 cases per 100,000 inhabitants for the study set and the validation set respectively. From 2014 to 2018, the average monthly incidence rate for all invasive tumours was similar to that observed in the validation set (Table 2). This table also shows that the average monthly incidence rate during the first year of COVID was significantly different from the training and validation sets. Only 32,457 new neoplasms were diagnosed that year resulting in an incidence rate of 54.77 cases per 100,000 people. However, the incidence rate during COVID's second year is similar to the time compressed between January 2014 and February 2020, with 71.03 instances per 100,000 persons on average.

Between January 2014 and December 2018, 207,768 new malignant neoplasms were registered in primary care, with little more than a half being men and almost half of the cases diagnosed in people who live in rural and the least deprived urban areas.

From Table 2 we can see that these percentages remain similar for the different sets.

From January 2019 until February 2022, the estimated rates of monthly new malignant neoplasm diagnoses were estimated using a time series linear regression model, the details and validation of which may be found in Annex C.

Figure 4 shows the observed and estimated rates of monthly new malignant neoplasm diagnosis, with 95% CI, since March 2020 until February 2022. For the validation period, observed incidences of malignant neoplasms were as expected as can be seen in

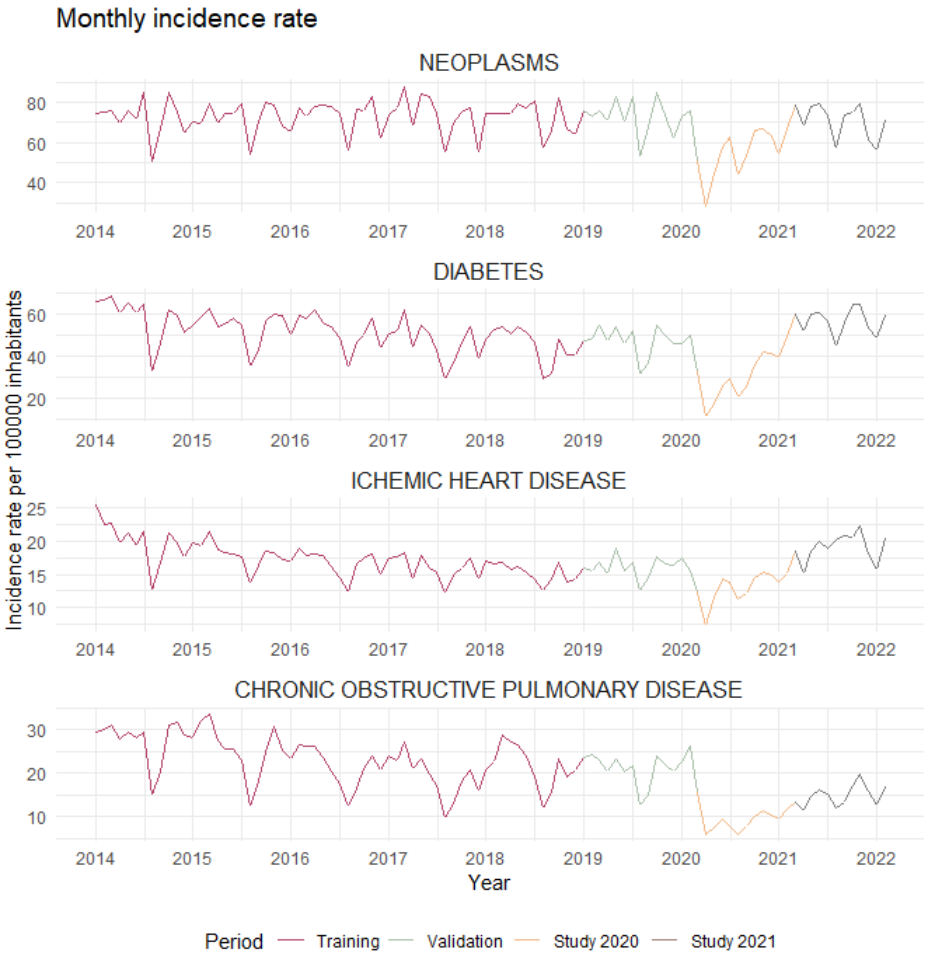


Figure 3: Incidence rate per 100,000 inhabitants for all the diseases with respect to the training, validation and study sets.

Appendix C. Since March 2020, observed incidences of malignant neoplasms have been significantly lower than expected for the entire population for three quarters of the COVID-19 period, as shown in Table 3. This period's incidence rates in November, December, and February are similar to the projected rate. Hence, during this period there was a lost of 30.61% (20.86, 38.22) of diagnoses. On the other hand, during the Second Year of COVID-19, the incidence was similar to the expected one for every month but January.

This fact can also be seen in a similar way per each of the stratums: sex, age and socioeconomic status as shown in Figures 5, 6, 7 respectively and also summarized in Table 3. We can observe in this table that the greatest decreases occurred during COVID-19's first year. Males in the sex stratum, adults up to 70 years old, and the 1st quartile of MEDEA (the most deprived ones) in the socioeconomic status stratum were the most affected, with only 16.7% of the incidence rate matching the predicted rate in all cases. On the other hand, the strata belonging to rural areas was the least affected during the first year of COVID-19. Here, 58.3% of the incidence rate matched our predictions.

Then, regarding the second year of COVID-19, the percentage of neoplasm diagnosis' recovery is close to 100% and the difference of the whole corresponds to the month of

Table 2: Number of Neoplasms with respect to each study set: training (2014-2018), validation set, first year of pandemic and second year of pandemic analysis sets and strats.

		Training set	Validation set	1st analysis set	2nd analysis set
TOTAL	Number of diagnoses	207,768	49,979	32,457	42,533
	Mean monthly incidence rate	72.81	72.98	54.77	71.03
		(70.69, 74.93)	(68.60, 77.35)	(47.21, 62.34)	(65.57, 76.48)
SEX	Women	96,123	23,701	15,642	20,988
		46.26%	47.42%	48.19%	49.35%
	Men	111,645	26,278	16,815	21545
		53.74%	52.58%	51.89%	50.65%
MEDEA	1st Q (least deprived)	48,986	11,630	7,151	9,598
		23.58%	23.27%	22.03%	22.57%
	2nd Q	32,011	7,700	5,025	6,503
		15.41%	15.41%	15.48%	15.29%
	3rd Q	40,772	9,756	6,320	8,520
		19.62%	19.52%	19.47%	20.03%
	4th Q (most deprived)	35,100	8,548	5,621	7,206
		16.89%	17.10%	17.32%	16.94%
	Rural	50,899	12,345	8,340	10,706
		24.50%	24.70%	25.70%	25.17%
AGE	Younger than 60	55,991	13,558	8,974	11,447
		26.95%	27.13%	27.65%	26.91%
	Between 60 and 70	54,691	12,797	8,173	10,453
		26.32%	25.60%	25.18%	24.57%
	Between 70 and 80	46,791	11,866	7,605	10,317
		22.52%	23.74%	23.43%	24.26%
	Older than 80	50,295	11,758	7,705	10,316
		24.21%	23.53%	23.74%	24.25%

January 2022 in every case.

Finally, we would like to comment on some particularities of the stratums. Regarding age stratum, Figure 6 shows that during the validation set all the observed incidence rates fall into the expected range but the group of people between 60 and 70 years old, where the observed incidence in June 2019 was lower than expected. For this stratum, during COVID-19's final year, the observed monthly incidence rate differed from the projections by non-statistically significant amounts. As a result, it was similar to pre-pandemic levels for all the second study period but for people younger than 80 during January 2022. Finally, from Appendix B, we can observe that during the validation set all the observed incidence rates fall into the predicted interval but the 3rd quartile of MEDEA, where the observed incidence in February 2019 was lower than expected. Regarding the second year of COVID-19, we can observe that, except in January 2022 for urban areas, the incidence rate for each month compressed in this period is similar to our predictions.

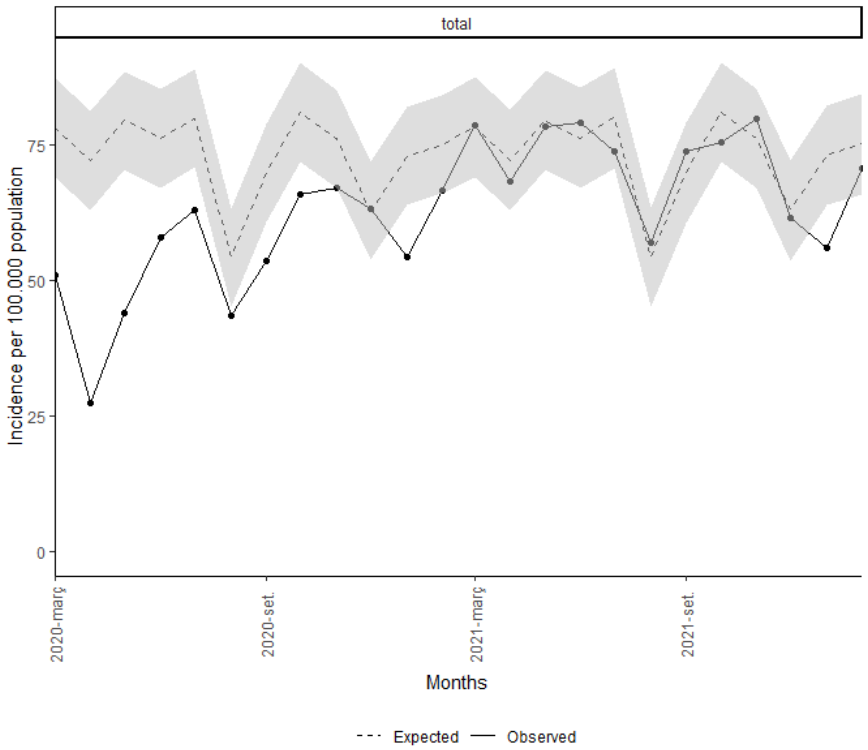


Figure 4: Predictions of total neoplasm's incidence rate per 100,000 inhabitants with 95% CI.

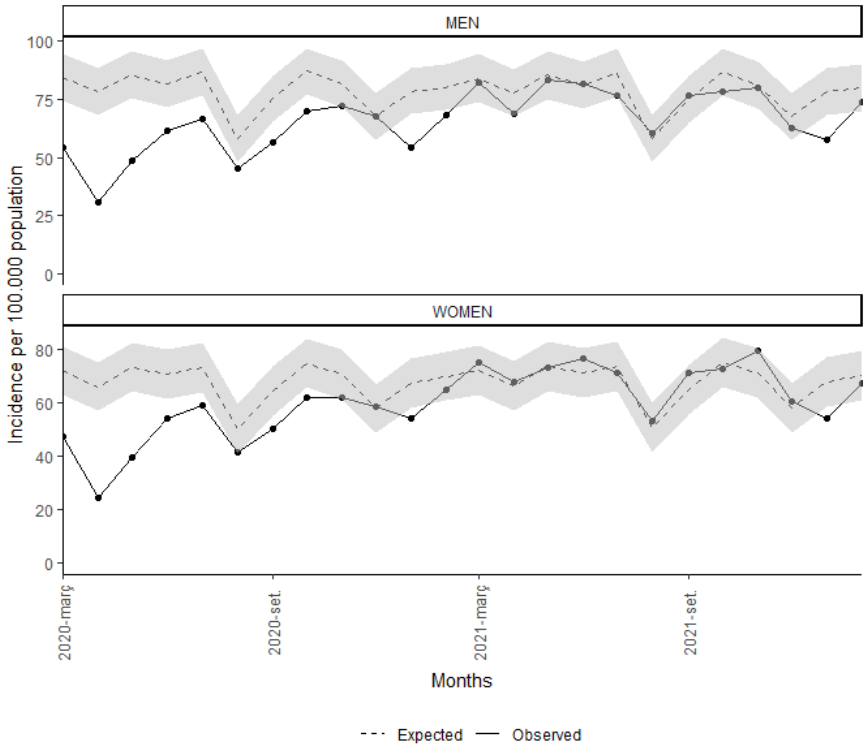


Figure 5: Predictions of total neoplasm's incidence rate per 100,000 inhabitants by sex with 95% CI.

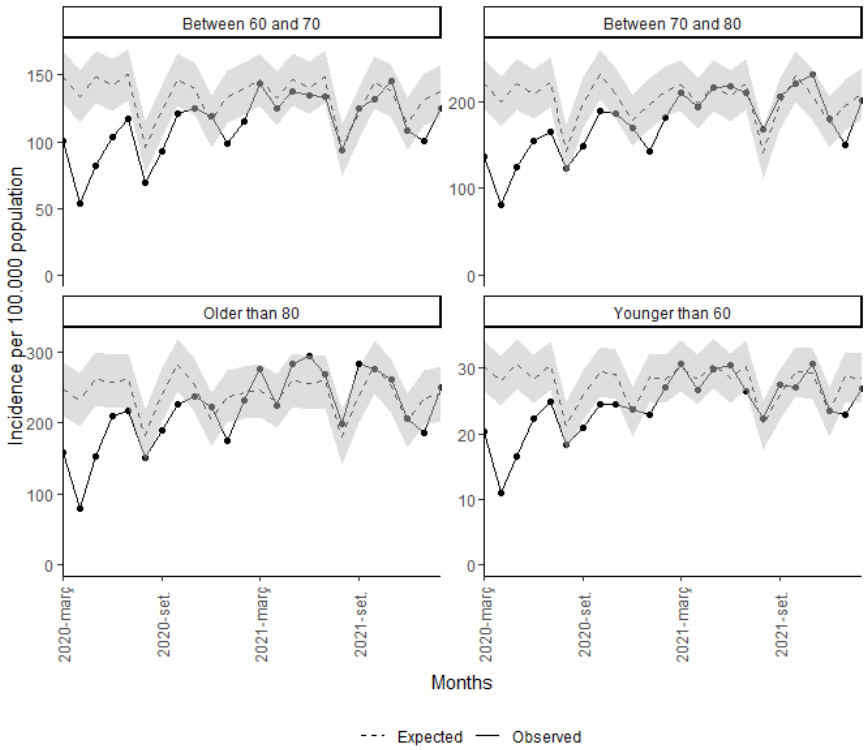


Figure 6: Predictions of total neoplasm's incidence rate per 100,000 inhabitants by age with 95% CI.

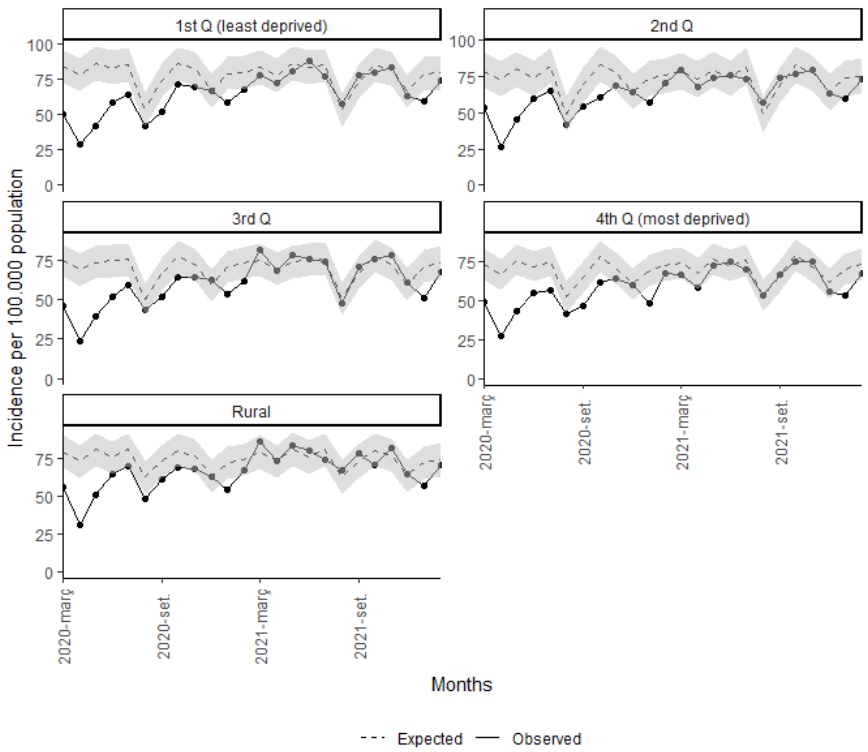


Figure 7: Predictions of total neoplasm's incidence rate per 100,000 inhabitants by socioeconomic status with 95% CI.

Table 3: Percentage of neoplasm diagnosis' recovery per study period and stratum in all sets.

	TOTAL	SEX		AGE				MEDEA				
		M	F	< 60	>=60 & <= 70	>70 & <= 80	> 80	1U	2U	3U	4U	Rural
Covid1	25	16.7	33.3	16.7	16.7	33.3	41.7	16.7	33.3	25	25	58.3
Covid2	91.7	91.7	91.7	100	91.7	91.7	91.7	91.7	91.7	91.7	91.7	100

3.2.2 Diabetes

Before the COVID-19 pandemic in Catalonia, from January 2014 to February 2020, 178,878 new diabetes cases were registered in primary care. This represents an average monthly incidence rate of 51.33 and 47.40 cases per 100,000 inhabitants for the study and the validation sets respectively. From 2014 to 2018, the average monthly incidence rate for all diabetes cases was not statistically different to that observed in the validation set (Table 4). However, we can observe that the mean from the two sets differs by 4 points. As we have seen in Figure 3, the diabetes incidence rate is globally higher during 2014 and 2015 than the other years included in the study set. In addition, during these first years we can observe a decreasing trend, although it remains more or less constant over 2016, 2017, 2018 and the validation set.

In order to be able to observe the aforementioned rate differences, all the monthly incidence rates have been depicted in Figure 8. According to this figure, we can see that 2014 rates are above those coming from the other training years. Therefore, we decided to exclude 2014 and 2015 from our training set in order to obtain a stable sample. In Figure 8, we can observe that once we have excluded them, 2019 rates are similar to the others from the training set. This behaviour can also be observed in Table 4. The mean monthly incidence rate from the study period is similar to the one from the validation set.

From the latter plot, Figure 10, we can empirically see that all monthly incidence rates from 2020 but the ones corresponding to January, February, November and December are lower than all the observed ones from the rest of the years. In addition, after April all the incidence rates are higher than the ones observed the years before in 2021. And finally, in 2022 the observed incidence rates are also higher than the ones observed in 2021, which shows a tendency to recover the diagnostic capacity of diabetes.

The latter fact can also be seen in Figure 9. To determine the means of each homogeneous section of the time series, a change point model is utilised. This figure confirms what we've seen before, in that the incidence rate for 2014 is higher than the remainder of the period, as well as 2015. We see that the time series is constant from the second half of 2016 to February 2020; then it declines abruptly in March and April 2020; then it increases in two steps, first at the end of 2020 and beginning of 2021; then later in 2021, it reaches a monthly incidence rate higher than the study period.

Additionally, in Table 4, we can also observe the fact that the monthly incidence rate during the first year of COVID was significantly different from the training and validation sets. Only 18,340 new diabetes cases were diagnosed this year resulting in an incidence rate of 30.95 cases per 100,000 people. Furthermore, the incidence rate during COVID's second year is neither comparable to the one from the time compressed between January 2016 and February 2020. Specifically, it is superior, with 33,860 instances per 100,000 persons on average which is an incidence rate of 56.53 per 100,000 inhabitants.

Between January 2016 and December 2018, 82,376 new diabetes cases were registered in primary care. A little more than half of these cases were diagnosed in men and almost a quarter in rural areas. Further details of the incidence distribution can be found

in Table 4 from where we can also see that this distribution remains similar during the different sets.

Table 4: Number of diabetes cases for every set: training (2014-2018), training (2016-2018), validation set, first year of pandemic analysis set and second year of pandemic analysis set and strats.

		2014-2018	2016-2018	Validation set	1st analysis set	2nd analysis set
Total	Number of diagnoses	146,418	82,376	32,460	18,340	33,860
	Mean monthly incidence rate	51.33	48.01	47.40	30.95	56.53
		(48.84, 53.83)	(45.13, 50.88)	(43.66, 51.13)	(23.85, 38.04)	(52.68, 60.38)
Sex	Women	63,182	35,650	13,939	8,075	15,364
		43.15%	43.28%	42.94%	44.03%	45.38%
	Men	83,236	46,726	18,521	10,265	18,496
		56.85%	56.72%	57.06%	55.97%	54.62%
MEDEA	1st Q (least deprived)	26,713	15,153	5,850	3,154	6,358
		18.24%	18.39%	18.02%	17.20%	18.78%
	2nd Q	22,853	12,713	4,991	2,655	5,056
		15.61%	15.43%	15.38%	14.48%	14.93%
	3rd Q	30,398	17,186	6,793	3,803	7,122
		20.76%	20.86%	20.93%	20.74%	21.03%
	4th Q (most deprived)	30,991	17,455	7,106	4,056	7,658
		21.17%	21.19%	21.89%	22.12%	22.62%
	Rural	35,463	19,869	7,720	4,672	7,666
		24.22%	24.12%	23.78%	25.47%	22.64%
AGE	Younger than 55	37,037	21,264	8,607	5,363	8,037
		25.30%	25.81%	26.52%	29.24%	23.74%
	Between 55 and 65	37,335	21,030	8,300	4,702	8,375
		25.50%	25.53%	25.57%	25.64%	24.73%
	Between 65 and 75	37,436	20,982	8,155	4,403	8,696
		25.57%	25.47%	25.12%	24.01%	25.68%
	Older than 75	34,610	19,100	7,398	3,872	8,752
		23.63%	23.19%	22.79%	21.11%	25.85%

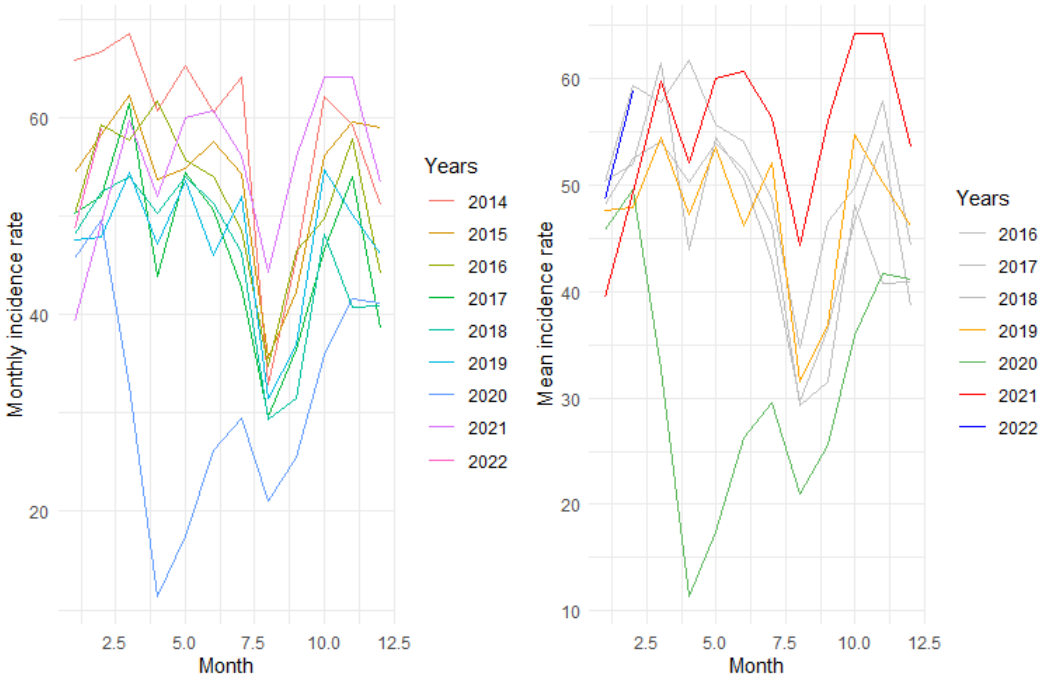


Figure 8: Monthly incidence rate comparison of diabetes for years compressed in the study period. On the left, we show the monthly incidence rate for the set between 2014 and 2022. On the right, we show the monthly incidence rate for the set between 2016 and 2022.

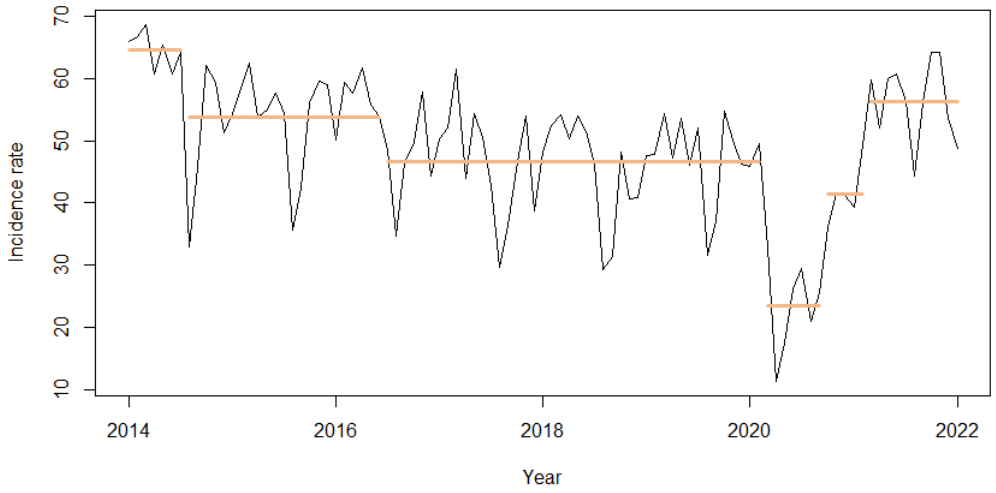


Figure 9: Mean of the diabetes incidence rate over time.

On the other hand, the results obtained from our model, a linear time series regression trained with data from 2016 to 2018, are shown below. All the specifications and the model validation can be found in Annex C.

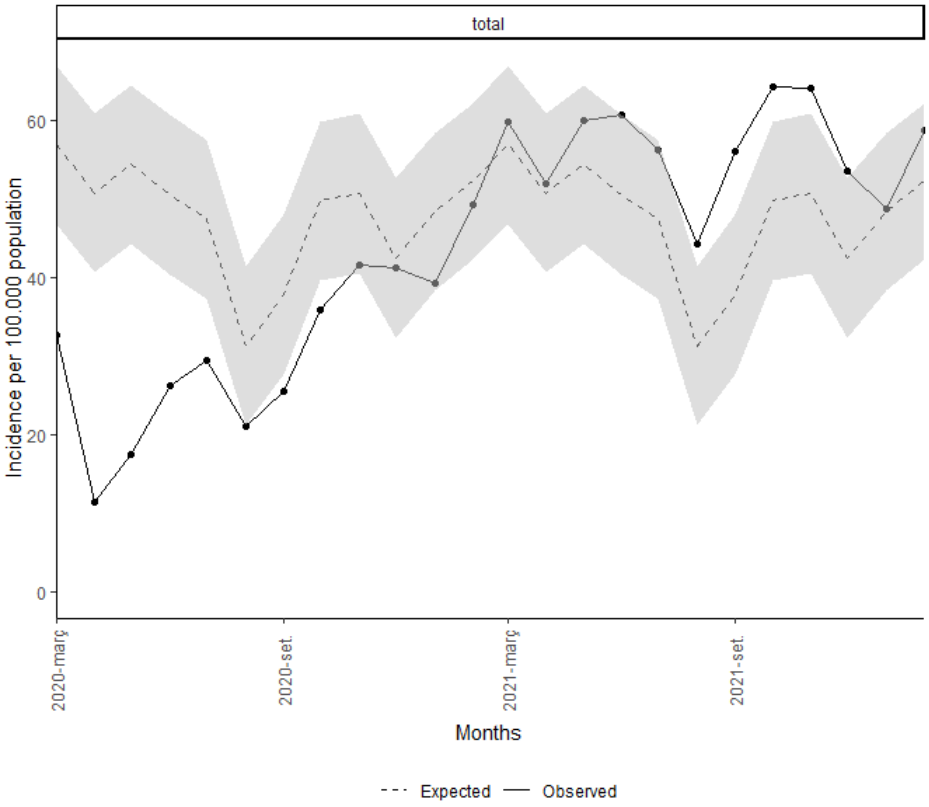


Figure 10: Predictions of total diabetes' incidence rate per 100,000 inhabitants with 95% CI.

From this global incidence rate plot, we can observe that since March 2020 and until November 2020 all the rates are lower than the expected range. From December 2021 until July 2021, the observed rates are similar to the predicted ones. During the second half of 2021, the detected incidence rate was greater than expected. And finally the incidence rate of the first two months of 2022 was similar to the expected one.

Note that the percentage of diabetes diagnosis' recovery is shown per stratum and period in Table 5. From here we can see that during the first year of COVID-19 in almost all the stratums the incidence rate of, approximately, 40% of the months were not in the expected range. Regarding the second year of the pandemic, on the other hand, the incidence rate of almost 60% of the months is similar to the expected one. Below, we are going to analyse each of the stratums.

Table 5: Percentage of diabetes diagnosis' recovery per study period and stratum in all sets.

	TOTAL	SEX		AGE				MEDEA				
		M	F	< 60	>=60 & <= 70	>70 & <= 80	> 80	1U	2U	3U	4U	Rural
Covid1	33.3	25	33.3	58.3	33.3	25	25	33.3	41.7	33.3	25	50
Covid2	58.3	75	41.7	91.7	83.3	50	25	58.3	91.7	50	33.3	91.7

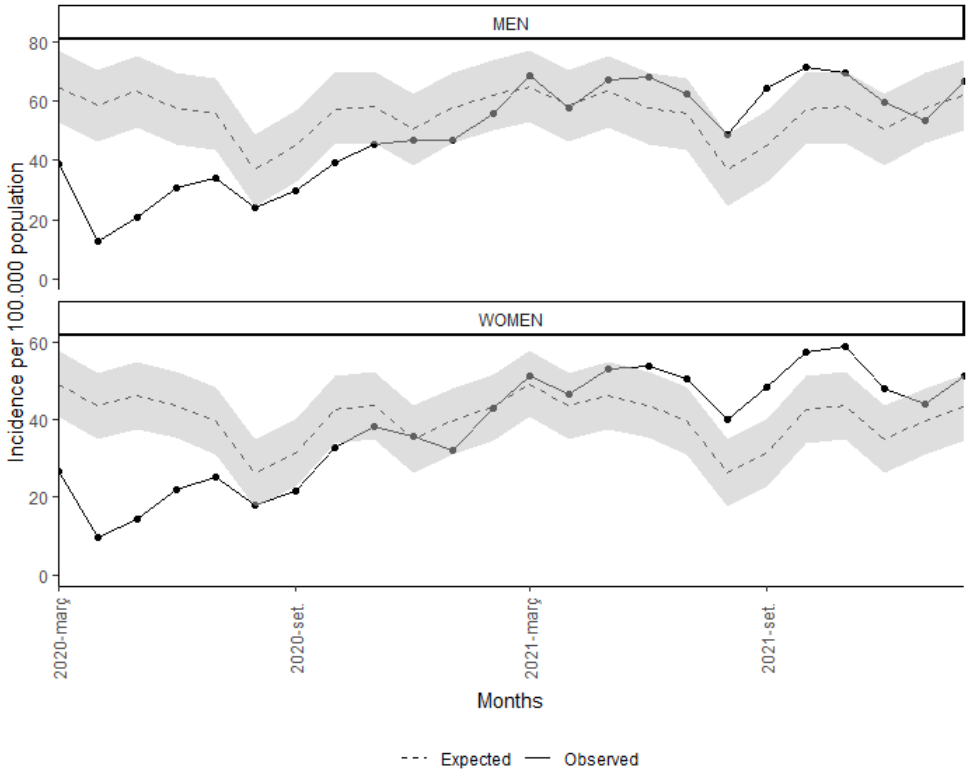


Figure 11: Predictions of total diabetes' incidence rate per 100,000 inhabitants by sex with 95% CI.

Regarding sex, from Table 4 we can conclude that during all the periods the percentage of cases in men is higher than for women. This fact can also be observed in Figure 11. From the latter plot we can also see that during the second half of 2021 the incidence rate was greater than the expected especially for women. While regarding men the incidence rate falls in the upper part of the interval, for women falls above the interval.

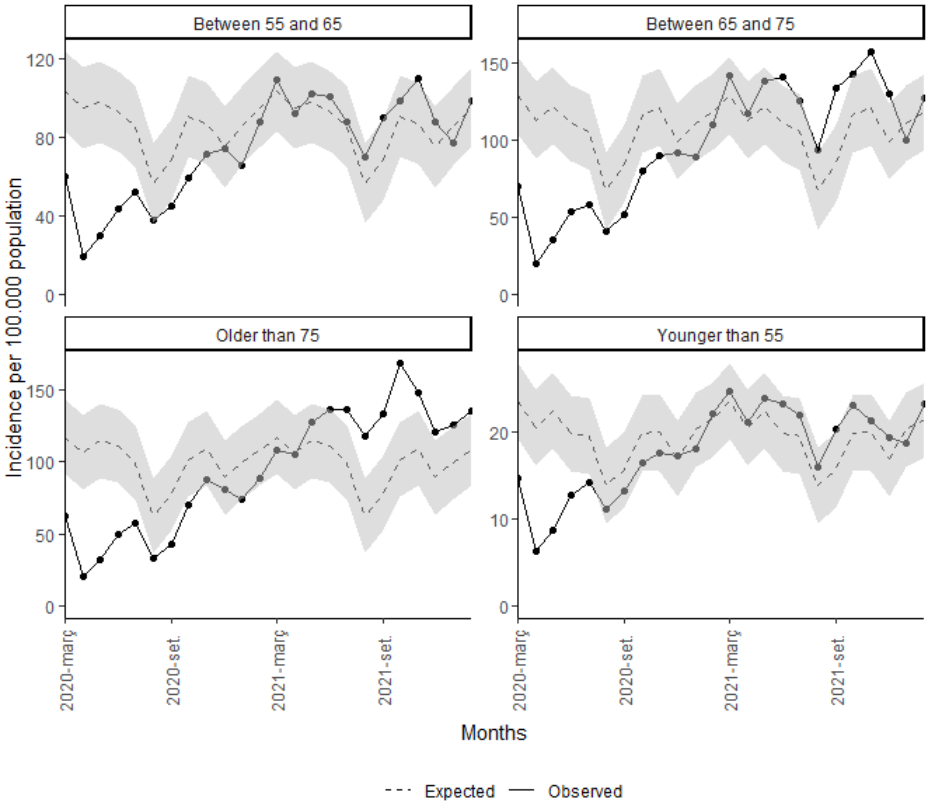


Figure 12: Predictions of total diabetes' incidence rate per 100,000 inhabitants by age with 95% CI.

Regarding age, from Table 4 we can see that the proportion of diagnoses for each age range is similar for the training, validation and the 2nd year of COVID-19 periods. However, these percentages vary slightly during the 1st year of COVID-19 sets. Percentages of diagnoses performed on people older than 75 and between 65 and 75 years old decreased by 2 and 1 point respectively, while the percentage of people younger than 55 increased by 4 points. This fact can also be observed in Figure 12. In the latter figure, we can see that the incidence rate of people younger than 65 years old is similar to the expected incidence during the last months of the first COVID-19's year and also during the second one. Nevertheless, during the same period of time and for people older than 65, the observed incidence rate falls in the upper part of the CI and between July 2021 and December 2021 the observed one is even higher than the expected one.

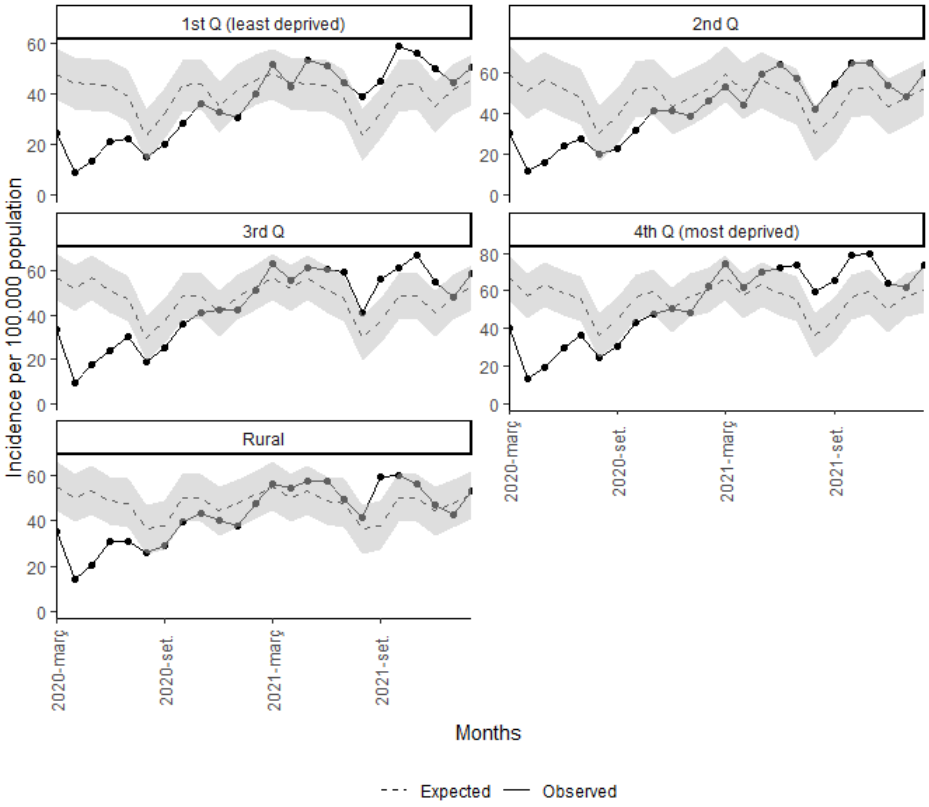


Figure 13: Predictions of total diabetes' incidence rate per 100,000 inhabitants by socioeconomic status with 95% CI.

Regarding socioeconomic status, in Table 4 we can observe that the percentage of incidence per each group remains similar over all the time periods. Figure 13 shows the decrease in incidence throughout the first half of COVID-19's first year. Following that, until June 2021, all incidence rates fell within the expected interval. Then, for all stratums except 2nd Q, the recorded incidences rise above the interval from June to December 2021. Furthermore, the 4th Q, which is the most impoverished, has a higher excess incidence.

As in several stratums and also in the global model we have observed that the incidence during specific months of the second year of COVID-19 is higher than the expected, the mean percentage of excess and underdiagnosis of each period is calculated. Therefore, the observed mean incidence per each study set is 30.9 and 56.5 respectively. The expected mean incidence per each study set is 48.0. Hence, during the first year of COVID-19 we lost 48.7% (33.09, 58.41) of the diagnoses while during the second year we detected 34.6% (5.39, 86.17) more cases. Taking into account the two study sets together, the observed monthly mean incidence rate is 43.7 per 100,000 inhabitants and the predicted one is 48.1.

3.2.3 Ischemic heart disease

Before the COVID-19 pandemic in Catalonia, from January 2014 to February 2020, 60,113 new ischemic heart disease cases were registered in primary care. This represents a mean monthly incidence rate of 17.2 cases per 100,000 inhabitants for the study set and an incidence rate of 16.11 cases per 100,000 individuals for the validation set. From 2014 to 2018, the mean monthly incidence rate for ischemic heart disease was comparable to that observed in the validation set (Table 6). This table also shows that the total mean monthly incidence rate during the first year of COVID was significantly lower than the training and validation sets. Only 7,742 new cases were diagnosed this year resulting in an incidence rate of 13.07 cases per 100,000 people. The incidence rate during COVID's second year is also superior to the time compressed between January 2014 and February 2020, with 19.08 instances per 100,000 persons on average, hence, higher than before the pandemic.

Table 6: Number of ischemic heart disease cases per each study set: training (2014-2018), validation set, first year of pandemic analysis set and second year of pandemic analysis set and strats.

		Training set	Validation set	1st analysis set	2nd analysis set
Total	Number of diagnoses	49,082	11,031	7,742	11,430
	Mean monthly incidence rate	17.20	16.11	13.07	19.08
		(16.51, 17.90)	(15.24, 16.97)	(11.62, 14.51)	(17.75, 20.40)
Sex	Women	16,435	3,671	2,520	3,925
		33.48%	33.28%	32.55%	34.34%
	Men	32,647	7,360	5,222	7,505
		66.52%	66.71%	67.45%	65.66%
MEDEA	1st Q (least deprived)	10,654	2,297	1,502	2,511
		21.71%	20.82%	19.40%	21.97%
	2nd Q	7,344	1,633	1,160	1,777
		14.56%	14.80%	14.98%	15.55%
	3rd Q	9,456	2,099	1,554	2,198
		19.27%	19.03%	20.07%	19.23%
	4th Q (most deprived)	9,327	2,199	1,568	2,326
		19.00%	19.93%	22.12%	20.35%
	Rural	12,301	2,803	1,958	2,618
		25.06%	25.41%	25.29%	22.90%
AGE	Younger than 60	12,673	2,864	2,182	2,668
		25.82%	25.96%	28.18%	23.34%
	Between 60 and 70	12,026	2,779	1,956	2,765
		24.50%	25.19%	25.26%	24.19%
	Between 70 and 80	12,218	2,923	1,951	3,147
		24.89%	26.50%	25.20%	27.53%
	Older than 80	12,165	2,465	1,653	2,850
		24.79%	22.35%	21.35%	24.93%

Between January 2014 and December 2018, 49,082 new ischemic heart disease cases were registered in primary care, with just one third of patients being women and a quarter belonging to rural areas. Further details of the distribution of cases among sex, socioeconomical status and age can be found in Table 6 where we can appreciate that these percentages remain similar during the different periods.

In order to determine whether the incidence rate during the two years of COVID is similar to the one before the pandemic, as shown in Annex III we tried to fit a time series regression but we could not validate the model. Therefore, we ran a Welch test to study the differences between means for each of the stratums previously defined. We specifically compared the years before the pandemic, the period between January 2014 until February 2020, with each year of COVID-19. The results from this test are displayed in Table 7. From this table, we can see that globally no the first year of COVID-19 monthly incidence rate either the one from the second year are similar to the rates before the pandemic in a 0.05 significance. While during the first year of the pandemic the incidence rate was statistically lower than the one observed before the COVID-19 outbreak, during the second year, it was statistically higher than the prepandemic one.

Regarding sex, the same pattern of the global rate was observed. Regarding socioeconomic status, we can observe that both COVID-19 years' incidence rates of ischemic heart disease are also lower than the prepandemic's ones showing the same pattern as the total incidence rate, lower during the first year and higher during the second one. However, regarding the rural area, we can see that the incidence rate during the second year is not statistically different to the prepandemic one at a 0.05 significance. Finally, in terms of age, during the first year of COVID-19, all the incidence rates were statistically lower from the incidences observed from 2014 until the beginning of 2020. However, during the second year, the incidence rate of people under 70 years old is statistically equal to the incidence rate before the pandemic, while the incidence rate of patients older than 70 years old is statistically higher than before the pandemic.

Table 7: Comparison of the monthly incidence rate (MIR) of ischemic heart disease cases before the pandemic and during each year of COVID-19 pandemic. Results from a Welch Two Sample t-test comparing each period with the one before the pandemic.

			January 2014 - February 2020	1st analysis set	2nd analysis set
TOTAL		MIR	17.00 (16.41, 17.59)	13.07 (11.62, 14.51)	19.08 (17.75, 20.40)
		p-value		5.373e-05	6.56e-03
SEX	Women	MIR	11.13 (10.68, 11.57)	8.32 (7.16, 9.48)	12.84 (11.74, 13.94)
		p-value		1.79e-04	6.63e-03
	Men	MIR	23.12 (22.37, 23.90)	18.03 (16.21, 19.85)	25.58 (23.79, 27.37)
		p-value		3.86e-05	1.47e-02
MEDEA	1st Q (least deprived)	MIR	17.07 (16.28, 17.86)	11.70 (10.05, 13.36)	19.34 (17.23, 21.44)
		p-value		6.35e-06	4.46e-02
	2nd Q	MIR	16.80 (16.14, 17.46)	12.83 (10.95, 14.06)	19.47 (17.71, 21.23)
		p-value		6.55e-04	7.44e-03
	3rd Q	MIR	15.60 (15.02, 16.17)	12.68 (11.29, 14.06)	17.76 (16.46, 19.06)
		p-value		6.55e-04	4.37e-03
	4th Q (most deprived)	MIR	17.63 (16.95, 18.32)	14.35 (12.33, 16.37)	21.17 (19.67, 22.67)
		p-value		4.63e-03	2.36e-04
	Rural	MIR	17.80 (17.12, 18.49)	13.80 (12.51, 15.08)	18.14 (18.09, 19.19)
		p-value		9.391e-06	0.576
AGE	Younger than 60	MIR	6.14 (5.94, 6.33)	5.21 (4.74, 5.67)	6.31 (6.02, 6.60)
		p-value		1.09e-03	0.305
	Between 60 and 70	MIR	33.77 (32.55, 34.99)	25.93 (22.53, 29.33)	35.95 (33.59, 38.30)
		p-value		2.93e-04	0.094
	Between 70 and 80	MIR	47.78 (45.66, 49.89)	34.27 (29.23, 39.32)	54.39 (49.24, 59.55)
		p-value		6.37e-05	0.020
	Older than 80	MIR	58.99 (55.97, 62.02)	40.21 (35.44, 44.97)	69.17 (61.37, 76.96)
		p-value		2.669e-07	0.018

3.2.4 Chronic obstructive pulmonary disease

Before the COVID-19 pandemic in Catalonia, from January 2014 to February 2020, 80,949 new chronic obstructive pulmonary disease cases were registered in primary care. This represents a mean monthly incidence rate of 23.23 cases per 100,000 inhabitants for the study set and an incidence rate of 21.45 cases per 100,000 individuals for the validation set. From 2014 to 2018, the mean monthly incidence rate for chronic obstructive pulmonary disease was similar to that observed in the validation set (Table 8). This table also shows that the total mean monthly incidence rate during the first year of COVID was significantly lower from the training and validation sets. Only 5,621 new cases were diagnosed this year resulting in an incidence rate of 9.48 cases per 100,000 people. The incidence rate during COVID's second year is also lower than the one before the pandemic, with 14.92 instances per 100,000 persons on average.

Between January 2014 and December 2018, 66,260 new chronic obstructive pulmonary disease cases were registered in primary care health centres, with less than a third of patients being women and more than a quarter living in the rural area. This information can be found in Table 8. In this table further details of the incidence distribution in terms of sex, age and socioeconomic status can be found and from here we can also see that distributions remain similar during the different study periods.

Observing the incidence rates before the pandemic, as shown in Figure 14, we have seen that during 2014 and 2015 the incidence rate was higher than the rest of monthly incidence rates observed in this period. Therefore, we decided to exclude 2014 and 2015 from the training set. Hence, the training set considered for this disease includes all the months between january 2016 and February 2020.

In Table 8, the information related to this period is depicted. As can be observed, 35,943 new chronic obstructive pulmonary cases were detected during this period, which leads to a mean monthly incidence rate of 20.94 cases per 100,000 people. The sex, age and socioeconomic distributions remain similar to the ones commented before.

Table 8: Number of chronic obstructive pulmonary disease cases per each study set: training (2014-2018), validation set, first year of pandemic analysis set and second year of pandemic analysis set and strats.

		2014-2018	2016-2018	Validation set	1st analysis set	2nd analysis set
TOTAL	Number of diagnoses	66,260	35,943	14,689	5,621	8,938
	Mean monthly incidence rate	23.23	20.94	21.45	9.48	14.92
		(21.76, 24.70)	(19.35, 22.53)	(19.35, 23.54)	(7.75, 11.22)	(13.37, 16.47)
SEX	Women	20,069	11,352	13,939	1,976	3,267
		30.29%	31.58%	34.13%	35.15%	36.55%
	Men	46,191	24,590	9,675	3,645	5,671
		69.71%	68.42%	65.87%	64.85%	63.45%
MEDEA	1st Q (least deprived)	13,474	7,349	3,023	1,093	1,946
		20.34%	20.42%	20.58%	19.44%	21.77%
	2nd Q	9,891	5,233	2,099	858	1,324
		14.93%	14.56%	14.29%	15.26%	14.81%
	3rd Q	13,021	7,091	2,818	1,086	1,665
		19.65%	19.73%	19.18%	19.32%	18.63%
	4th Q (most deprived)	13,062	7,031	2,974	1,072	1,925
		19.71%	19.56%	20.25%	19.07%	21.54%
	Rural	16,812	9,247	3,775	1,512	2,078
		25.37%	25.72%	25.70%	26.90%	23.25%
AGE	Younger than 60	18,200	10,118	4,140	1,572	2,321
		27.47%	28.15%	28.18%	27.97%	25.97%
	Between 60 and 70	18,972	10,400	4,264	1,455	2,408
		28.64%	28.94%	29.03%	25.89%	26.94%
	Between 70 and 80	18,135	9,507	3,928	1,513	2,448
		27.37%	26.45%	26.74%	26.92%	27.39%
	Older than 80	10,953	5,917	2,357	1,081	1,761
		16.53%	16.46%	16.05%	19.23%	19.79%

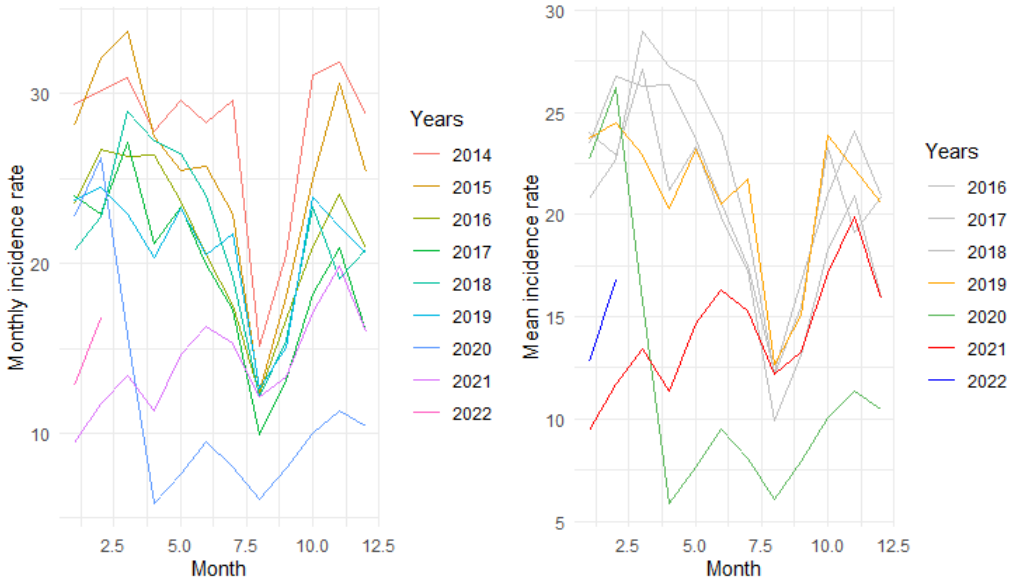


Figure 14: Monthly incidence rate comparison of chronic obstructive pulmonary disease for years compressed in the study period. Left 2014-2022, right 2016-2022, the considered study period.

Figure 15 also allows us to appreciate that after March 2020 the observed incidence rate differs a lot from the expected one fed with historical data. We can also observe that during the first part of 2021 it was also lower than the expected range. Since June and until December 2021 the incidence rate was similar to the one from 2017. And finally, during 2022 monthly incidence rates were greater than the observed during the same period during 2021 but below all the monthly incidence rates observed in the years before the pandemic. Therefore, the incidence rate during the pandemic years was much lower than the previous ones.

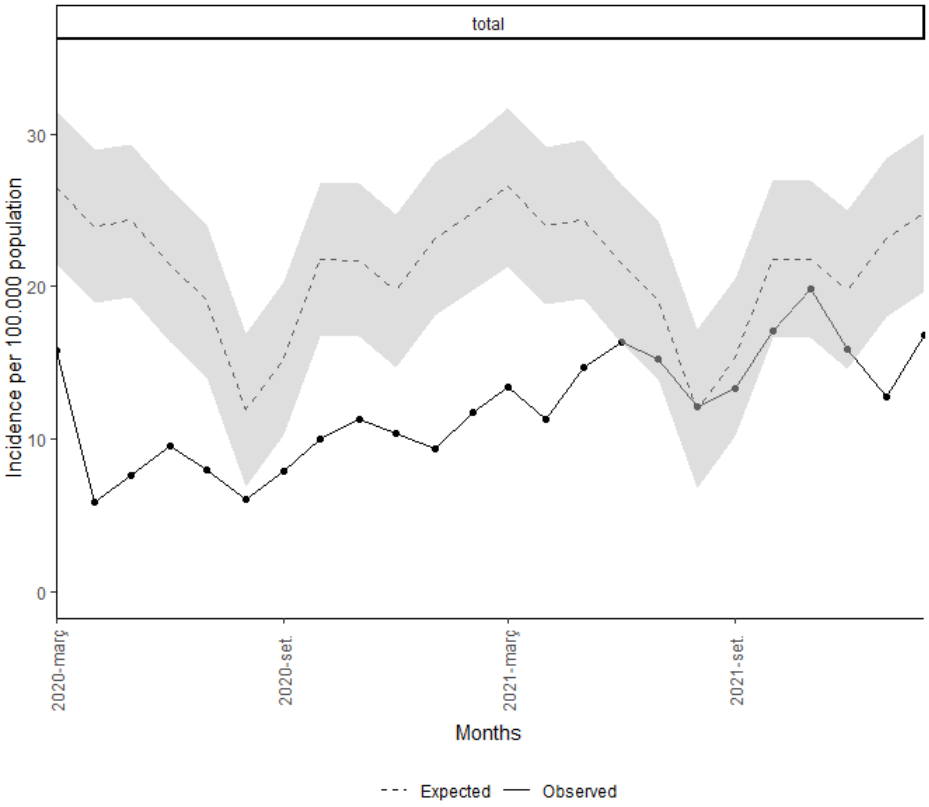


Figure 15: Monthly incidence rate comparison of chronic obstructive pulmonary disease for years compressed in the study period. Left 2014-2022, right 2016-2022, the considered study period.

In Figures 15, 16, 17 and 18 observed and estimated (with a 95% CI) incidences are shown. Note that all the specifications and the model validation can be found in Annex C. From these figures we may appreciate that the trained model fits accurately the data observed during 2019. Nonetheless, we can see that regarding men there was a drop in the incidence rate in February and March 2019. In addition there are some monthly incidence rates for other stratums that are not in the predicted range; for instance, April 2019 for people between 60 and 70 years old and also March, April and May 2019 for the 3rd quartile of MEDEA population and March 2019 for the 2nd quartile of MEDEA are below the expected range. On the other hand, during July 2019 for the least deprived level of the socioeconomic status and from July to October 2019 for the 2nd quartile of MEDEA, the observed monthly incidence rates are higher than the expected.

On the other hand, we can see that it abruptly drops in March 2020 but even more in April 2020 for all stratums. After that month the incidence rate remains lower than expected for all the stratums too but people older than 70 years old. From this age range, the incidence rate from some months during 2020 is similar to the predicted one.

Finally, regarding the last year of COVID-19, we can observe that its incidence rate globally is lower than the expected one or similar but falling to the lower part of the predicted range. However, the behaviour with respect to different stratums may change. Regarding sex, we can observe that the incidence rate for women is more similar to the

expected one than for men as can be seen in Table 16. However, both percentages of recovery are quite low: 58.3%, 41.7% of men and women respectively that are similar to the incidence rates before the pandemic. On the other hand, regarding age stratums we can see that the group of people older than 80 years old is the one that has better recovered the incidence rate. Apart from being the group with a higher percentage of observations similar to the predictions, it is the one that the observed rates belong to the higher part of the CI. The less recovered age groups are those with people under 70 years old. Finally, regarding socioeconomic status and also from this period of time, we can observe from Figure 18 and Table 9 we can see that the area with less percentage of recovery is the 3rd MEDEA quartile. It is also important to note that although for 1st, 2nd and 4th quartiles of MEDEA and in addition for the rural population there is a recovery during the last months of 2021, for all of them in January 2022 the observed incidence rate was lower than the expected one.

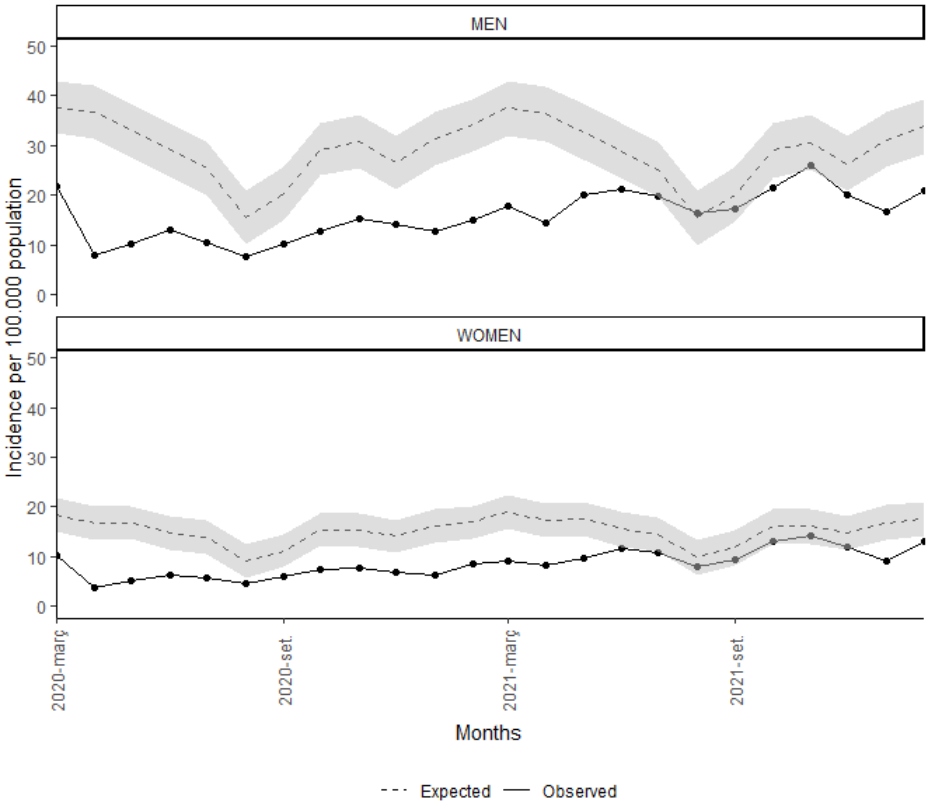


Figure 16: Chronic obstructive pulmonary disease's incidence per 100,000 inhabitants per sex.

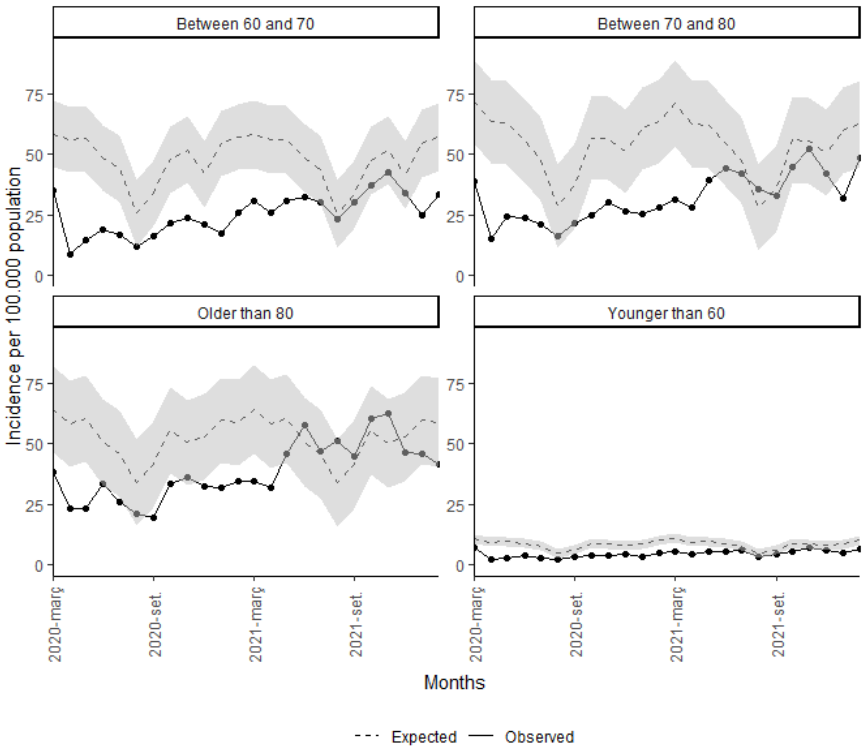


Figure 17: Chronic obstructive pulmonary disease's incidence per 100,000 inhabitants per age.

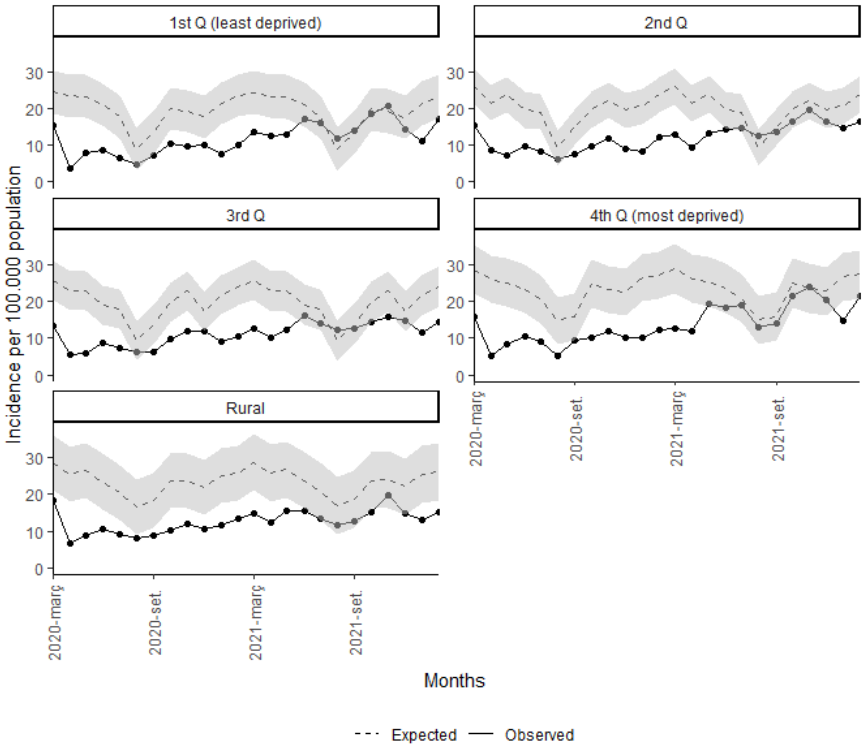


Figure 18: Chronic obstructive pulmonary disease's incidence per 100,000 inhabitants per socioeconomic status.

As in several stratums and also in the global model we have observed that the incidence during specific months of the first and second years of COVID-19 is lower than the expected, the mean percentage of underdiagnosis of each period is calculated. Therefore, the observed mean incidence per each study set is 9.48 and 14.92 respectively. The expected mean incidence per each study set is 21.0. Hence, during the first year of COVID-19 we lost 55.10% (41.16, 63.70) of the diagnoses while during the second year we lost 43.83% (28.98, 53.59) more cases.

Table 9: Percentage of diabetes diagnosis' recovery per study period and stratum in all sets.

	TOTAL	SEX		AGE				MEDEA				
		M	F	< 60	>=60 & <= 70	>70 & <= 80	> 80	1U	2U	3U	4U	Rural
Covid1	8.3	0	0	0	0	16.7	25	25	8.3	0	8.3	0
Covid2	58.3	58.3	41.7	41.7	58.3	66.7	75	66.7	50	25	75	58.3

4 Discussion

In this section we summarise the key findings of our study, provide interpretation and place them in the context of our research. We discuss the implications and acknowledge the limitations of the conducted study, while providing a brief look at potential follow-up research studies. Finally we conclude with a restatement of the most significant findings and their implications.

Our study analyses the incidence of four main chronic diseases: neoplasms, diabetes, ischemic heart disease and chronic obstructive pulmonary disease during the COVID-19 pandemic. We place special emphasis on the second year of the pandemic, to see if primary care has recovered the diagnostic capability to similar levels prior to the pandemic.

We compare the observed data for every disease with the predicted output by our time series linear regression models fitted with historical data from 2014-2019. We show that the above-mentioned diseases do not exhibit the same level of recovery between them. With respect to our model predictions and analysis, primary health care has maintained a similar level of diagnosis for neoplasms. Diabetes and ischemic heart disease however, show an increase in the incidence rate. Conversely, chronic obstructive pulmonary diseases decrease the rate of occurrence.

With regard to neoplasms, we have observed that they maintain a similar incidence rate during all the years of study before the pandemic, with a stable trend and a very marked seasonality. Thus, after validation our models have been re-trained with data between 2014 and 2019.

Notably, during the second year of COVID, the incidence rate has recovered from pre-pandemic levels in 91.7% of the months, i.e. 11 out of 12 months in total. The month that prevents a 100% recovery level is January, where the rate fell below the expected range and recovered next month in February. The latter finding however, didn't affect people under the age of 60 or those living in rural areas.

In January 2022 Catalonia experienced the peak of the sixth wave of COVID-19. During this period, as shown on the Generalitat de Catalunya's pandemic monitoring website [39], more than 100,000 daily visits to primary health care centres were recorded, with the main reason being the COVID-19. This led to overcrowding in primary care and in some cases perhaps fewer visits to patients with other pathologies, as there was very little availability of visits.

On the other hand, it should be noted that although in general during the second year of COVID we are at similar levels to pre-pandemic levels. It may be that many of the missed diagnoses will never be recovered because of the excess mortality rate in Spain during that period, as shown by the daily Mortality Monitoring (MoMo) system [40].

In addition, as diagnoses that should have been detected earlier are still expected to be recovered, it could be that all these delays lead to a worse prognosis for patients if the disease is detected in advanced stages [41, 42, 43]. Patients whose diagnosis was delayed during the COVID-19 epidemic have a new and significant issue for the

health-care system: prompt treatment and quality support. According to models based on overall cancer diagnoses in the UK, a 2-month delay in detection for 50% of individuals diagnosed with stage I-III cancer might result in a 6% increase in fatalities within 10 years [44]. Delayed diagnosis is likely to have a meaningful impact on patient morbidity and mortality, and is especially worrying for malignancies that can (quickly) grow to a more advanced state before being identified. The burden on patients and oncological care systems is increasing as a result of the move toward higher stage tumours at diagnosis. These high-stage cancers may require more intensive treatment. Excess burden could increase treatment delays and negatively effect the patient's prognosis if oncological care capacity is surpassed. Therefore, in conclusion, in regards to neoplasms, the diagnosis capacity during the second year of the pandemic is similar to the one we had before.

Regarding diabetes, we decided to exclude years 2014 and 2015 from the training period since the incidence rate observed in these years is higher than the rest of the training period. This could be due to the fact that in 2014 indicators of diagnostic adequacy appeared [45]. These indicators could lead to a better registering of diabetes, since they check that if someone has diabetes characteristics, for example, laboratory results indicating that the patient has diabetes, then he/she needs to have the diagnosis. These indicators also check the quality of diagnosis the other way around. In other words, they check that if someone has a diabetes diagnosis he/she has the characteristics that confirm it. We believe the above reasoning to be a key indicator to justify the increase of diagnoses during this two years period. Therefore, we decided to train our model with data from 2016 to 2019, excluding 2014 and 2015 from the training set.

We have observed that the time series of diabetes shows an overall decreasing trend. However, we can see that during 2016 and 2017 this decreasing rate is a bit higher than 2018 and 2019 when it flattens. Since there are some changes of trend during the study period, and since we are not sure whether the trend is real or is residual from what we have observed and discussed before, we have decided to exclude it from our model. Consequently, the regression time series linear model used in this case is based on the baseline and seasonalities. By excluding the trend, we can see that the model meets its assumptions and is adequately fitted in the validation period.

During the second year of COVID-19, we discovered that there were more diabetes incident diagnoses than predicted based on previous historical evolution. Women, those over the age of 75, and those in the fourth quartile of MEDEA's socioeconomic position, i.e. the most deprived ones, had the highest rate of recovery. These data suggest that the Catalan primary health care system has statistically higher diagnostic capacity than predicted. The diagnostic excess shown in the second year of the pandemic, however, does not compensate for the COVID-19 reduction seen in the first year where a lot of diagnoses were missed.

Based on this research we cannot infer that we have diagnosed those who were not diagnosed with diabetes during the first year of COVID-19. Furthermore, we cannot establish whether the current excess relates to those not recognized during the first year of the pandemic, or simply that we are detecting more diabetes because the percentage of population with diabetes has grown.

In terms of ischemic heart disease, we attempted to fit the trend by modelling the incidence rate time series using a linear time series regression model with two differentiations. The model assumptions, however, could not be validated since there is a connection between the residuals and the latter's distribution does not follow a normal distribution, as shown in Appendix C. Therefore, we opted to use a Welch Two Sample t-test in order to compare the incidence rates between the study periods.

We found that during the first year of COVID-19 there were statistically significantly fewer cases of ischemic heart disease diagnosed. During the second year, however, there was an excess of diagnoses. In terms of the defined stratums, the latter fact holds true for all of them but for people under 70 years old and rural population. For these groups, the observed incidence rates are similar to those before the pandemic, not superior.

Finally, we discovered that the incidence rate of chronic obstructive pulmonary disease is not comparable to that before the pandemic, but rather lower.

The incidence rate during 2014 and 2015 was higher than the rest from the study period. In 2014, similar to diabetes, indicators of diagnostic appropriateness for chronic obstructive pulmonary disease appeared [45]. Therefore, we also opted to exclude these years from the study period. Thus, we trained our models with data just from 2016 to 2019. Aside from this peculiarity, the incidence rate of chronic obstructive pulmonary disease has another singularity. During 2017 the incidence rate was lower than the one observed during the rest of the years, yielding a parabolic trend. We considered it to be an outlier, since there was no remarkable reason for which it may have happened. Accordingly, we decided to clean the data as stated in methods in order to fit the model including this year as shown in Appendix C.

We can conclude from our results that the overall incidence rate is lower than the one expected by our models fed with pre-pandemic data. During the first year of COVID-19 just 8.3% of the monthly average incidence rate was similar to the expected. During the second year of pandemic 58.3% of the observations are similar to the expected ones estimated with prepandemic data. However, 41.7% of the measurements are still lower than expected. Therefore, for this particular disease the Catalan primary care health system has neither recovered the diagnosis capacity that it had before the pandemic nor recovered the diagnoses that were lost during the COVID-19 outbreak. However, we have observed that during the last months the observed incidence rate falls in the lower part of the predicted IC95%. Hence, by the end of the year a diagnoses recovery could be observed and further analysis would be required.

Spirometry is one of the most commonly used diagnostic tests to identify chronic obstructive pulmonary disease. During the pandemic different institutions recommended limiting pulmonary function tests, exclusively to indispensable cases [46, 47, 48] to contain hospital spread of Sars-Cov-2 as these tests imply to blow and generate aerosols that could spread the virus. Hence, all these policies may have reduced the number of tests performed as can be seen in Appendix D and consequently lowered the number of diagnoses being detected. Hence, some strategy should be implemented in order to appropriately diagnose this pathology.

There are many papers that have analysed the drop of incident diagnoses and the effect of COVID-19 in chronic diseases [17, 49, 44, 50] and many others that propose possible policies and solutions to overcome the pandemic effects on the detection, prevention and intervention of chronic diseases [51]. Some are studies performed in the same region for different periods and obtain results similar to the ones we have found (20). However, we have found little literature regarding incidence recovery of chronic diseases and just for neoplasms. Specifically, in Belgium a study demonstrated that in the end of 2020 the decline in diagnoses was only partially and variably recovered [52]. These results are similar to the ones we have obtained, regarding neoplasms. On the other hand, we have found that in New-Zeland, the elimination strategy used to combat the COVID-19 pandemic has saved the population and the health system that serves it from the pandemic's worst effects. COVID-19 had a less dramatic influence on New Zealand's cancer care system. There was active national planning to guarantee that disruptions were actively handled. As a result, while there were significant drops in cancer registrations and key diagnostic services during the national closure in late March and April 2020, these services have since restored to near-normal levels showing no evidence of ongoing disruptions in cancer registration, diagnostic services, or treatment [53]. Therefore, directly comparing our results to those of the latter could be highly misleading if we do not take into account the radically different measures taken by the respective health authorities. Besides the above-mentioned studies, we have not been able to find any other research related to any other chronic disease but neoplasms.

4.1 Strengths and limitations

This study does not go without limitations and in this section we are going to analyse them as well as the strengths of the project.

First of all, the main limitation that faces this project is that the data used comes from the electronic health data of primary health care. To mitigate this limitation, we have validated the data quality and it's worth noting that this data is of excellent quality and includes around five million people, or 75% of Catalans. Data from Catalan electronic health records has been validated in several studies [**34**]. Nonetheless, we did not verify that the data was correct for each patient.

Furthermore, because we only utilise data from primary health care, we could be losing certain diagnoses that may be performed in a hospital. Nonetheless, these are sure to recover at some point in the future, since these diseases are primarily controlled by primary care practices, and patients will eventually go to their GP.

As strengths, we would like to remark that our estimation used data from 5 years and we validated our methods in a year not affected by the COVID-19 pandemic. In addition, it is worth noting that this is a pioneer study since we have found very few literature talking about the incidence recovery of chronic diseases. As a matter of fact, just research about neoplasms has been found globally. To the best of our knowledge our study is the first one analysing the incidence recovery of diabetes, ischemic heart disease and chronic obstructive pulmonary disease worldwide. Furthermore, there is no bibliography on the

incidence recovery in Catalonia. Hence, it is the first study analysing the recovery of chronic diseases' incidence two years after the COVID-19 breakdown.

4.2 Conclusions

In conclusion, our study shows a reduction in registered neoplasms, diabetes, ischemic heart disease and chronic obstructive pulmonary disease incidence during the first year of COVID-19 pandemic. This suggests that there are diagnoses that are being delayed, which may lead to negative health outcomes.

During the second year of the pandemic, the studied diseases exhibit different levels of recovery between them. While neoplasms have achieved a similar level of diagnoses, diabetes and ischemic heart disease show an increase in the incidence rate while chronic obstructive pulmonary disease shows a decrease in the rate of occurrence. Hence, for all the studied diagnoses but chronic obstructive pulmonary disease, the diagnosis capacity has been recovered to levels similar, for neoplasms, or higher, for diabetes and ischemic heart disease, than before the pandemic.

Furthermore, some policies should be adopted in order to counteract the ones taken during the outbreak of the pandemic, as for example the ones related to spirometries.

In addition, long-term studies should be performed in order to evaluate the future effects and the consequences of this situation, the impact of the pandemic and the infra diagnostic rate.

5 Glossary

- CI: Confidence interval
- COVID-19: Coronavirus disease 2019
- EHR: Electronic health record
- ICD-10: International Classification of Diseases 10th revision
- ICS: Institut Català de la Salut (Catalan Institute of Health)
- MEDEA: Mortality in small areas of spain & socioeconomic & environmental inequalities
- NCD: Non-communicabe diseases
- NPI: Non-pharmaceutical interventions
- Sars-Cov-2Severe Acute Respiratory Syndrome Coronavirus 2
- SISAP: Serveis de la Informació del Sistema d'Atenció Primària (Primary Care Information System)
- WHO: World Health Organization

References

[1] *COVID-19 Map*. https://coronavirus.jhu.edu/map.html. Accessed: 2022-05-22.

[2] Bernard Stoecklin S Rolland P Silue Y et al. “First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020.” In: (). DOI: 10.2807/1560-7917..

[3] et al. C. Sohrabi Z. Alsafi. “World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19).” In: (). DOI: 10.1016/j.ijsu.2020.02.034..

[4] Maria Nicola et al. “The socio-economic implications of the coronavirus pandemic (COVID-19): A review”. In: *International Journal of Surgery* 78 (2020), pp. 185–193. ISSN: 1743-9191. DOI: https://doi.org/10.1016/j.ijsu.2020.04.018. URL: https://www.sciencedirect.com/science/article/pii/S1743919120303162.

[5] Renyi Zhang et al. “Identifying airborne transmission as the dominant route for the spread of COVID-19”. In: *Proceedings of the National Academy of Sciences* 117.26 (2020), pp. 14857–14863. ISSN: 0027-8424. DOI: 10.1073/pnas.2009637117. eprint: https://www.pnas.org/content/117/26/14857.full.pdf. URL: https://www.pnas.org/content/117/26/14857.

[6] Birkmeyer J.D. Barnato A. Birkmeyer N. Bessler R. Skinner J. “The Impact of the COVID-19 Pandemic on Hospital Admissions in the United States.” In: ().

[7] Bhatt A.S.; Moscone A.; McElrath E.E.; Varshney A.S.; Claggett B.L.; Bhatt D.L.; Januzzi J.L.; Butler J.; Adler D.S.; Solomon S.D.; et al. “Fewer Hospitalizations for Acute Cardiovascular Conditions During the COVID-19 Pandemic.” In: (2020).

[8] Sarac N.J.; Sarac B.A.; Schoenbrunner A.R.; Janis J.E.; Harrison R.K.; Phieffer L.S.; Quatman C.E.; LV T.V. “A Review of State Guidelines for Elective Orthopaedic Procedures During the COVID-19 Outbreak.” In: (2020).

[9] Hartnett K.P.; Kite-Powell A.; DeVies J.; Coletta M.A.; Boehmer T.K.; Adjemian J.; Gundlapalli A.V.; “National Syndromic Surveillance Program Community of Practice. Impact of the COVID-19 Pandemic on Emergency Department Visits—United States”. In: (2020).

[10] Ferguson NM Laydon D Nedjati-Gilani G et al. “Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand.” In: (2020). DOI: https://doi.org/10.25561/77482..

[11] Ayouni I. Maatoug J. Dhouib W. et al. “Effective public health measures to mitigate the spread of COVID-19: a systematic review.” In: (2021). DOI: https://doi.org/10.1186/s12889-021-11111-1.

[12] *Boletín oficial del estado (BOE). Real Decreto 463/2020, de 14 marzo, por el que se declara el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19.* https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-3692. Accessed: 2022-01-19.

[13] J. Robitscher. L. Ruth J. Alongi. “Confronting The Health Debt: The Impact Of COVID-19 On Chronic Disease Prevention And Management.” In: (2021). DOI: 10.1377/hblog20210914.220940.

[14] Anteneh D. Dandena F Teklewold B. “Impact of COVID-19 and mitigation plans on essential health services: institutional experience of a hospital in Ethiopia.” In: (Oct. 2021). DOI: 10.1186/s12913-021-07106-8..

[15] Bernhard Michalowsky et al. “Effect of the COVID-19 lockdown on disease recognition and utilisation of healthcare services in the older population in Germany: a cross-sectional study”. In: *Age and Ageing* 50.2 (Nov. 2020), pp. 317–325. ISSN: 0002-0729. DOI: 10.1093/ageing/afaa260. eprint: https://academic.oup.com/ageing/article-pdf/50/2/317/40967767/afaa260.pdf. URL: https://doi.org/10.1093/ageing/afaa260.

[16] Yogini Chudasama et al. “Impact of COVID-19 on routine care for chronic diseases: A global survey of views from healthcare professionals”. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (June 2020). DOI: 10.1016/j.dsx.2020.06.042.

[17] Héctor Pifarré i Arolas et al. “Missing Diagnoses during the COVID-19 Pandemic: A Year in Review”. In: *International Journal of Environmental Research and Public Health* 18.10 (2021). ISSN: 1660-4601. DOI: 10.3390/ijerph18105335. URL: https://www.mdpi.com/1660-4601/18/10/5335.

[18] Rick F Odoke W van den Hombergh J Benzaken AS Avelino-Silva. “Impact of coronavirus disease (COVID-19) on HIV testing and care provision across four continents.” In: *HIV Med.* (Oct. 2021). DOI: 10.1111/hiv.13180.

[19] Rymaszewska A.Acta Parasitol. Piotrowski M. “The Impact of a Pandemic COVID-19 on the Incidence of Borreliosis in Poland.” In: (Jan. 2022). DOI: 10.1007/s11686-021-00495-0.

[20] World Health Organization. *Noncommunicable diseases country profiles 2018*. 2018.

[21] Hans Henri P Kluge et al. “Prevention and control of non-communicable diseases in the COVID-19 response”. In: *The Lancet* 395.10238 (2020), pp. 1678–1680. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(20)31067-9. URL: https://www.sciencedirect.com/science/article/pii/S0140673620310679.

[22] World Health Organization. *COVID-19 significantly impacts health services for non-communicable diseases*. 2020.

[23] Chang AY Cullen MR Harrington RA Barry M. “The impact of novel coronavirus COVID-19 on noncommunicable disease patients and health systems: a review.” In: *J Intern Med.* 289.42 ().

[24] Debra Patt et al. “Impact of COVID-19 on Cancer Care: How the Pandemic Is Delaying Cancer Diagnosis and Treatment for American Seniors”. In: *JCO Clinical Cancer Informatics* 4 (2020), pp. 1059–1071. DOI: 10.1200/CCI.20.00134.

[25] C.H. Earnshaw H.J.A Hunter E. McMullen C.E.M Griffiths and R.B Warren. “Reduction in skin cancer diagnosis, and overall cancer referrals, during the COVID-19 pandemic.” In: *Br J Dermatol* 183 (2020), pp. 792–794. DOI: 10.1111/bjd.19267.

[26] K. Podwojcic M. Maluchnik and B. Wieckowska. “Decreasing access to cancer diagnosis and treatment during the COVID-19 pandemic in Poland.” In: *Acta Oncol.* 60 (2021), pp. 28–61. DOI: 10.1080/0284186X.2020.1837392.

[27] H.M. Peacock T. Tambuyzer F. Verdoodt F. Calay H.A. Poirel H. De Schutter J. Franicart N. Van Damme L. Van Eycken. “Decline and incomplete recovery in cancer diagnoses during the COVID-19 pandemic in Belgium: a year-long, population-level analysis.” In: *ESMO Open.* 60 (2021). DOI: https://doi.org/10.1016/j.esmoop.2021.100197.

[28] Coma E Guiriguet C Mora N Marzo-Castillejo M Benítez M Méndez-Boo L Fina F Fàbregas M Mercadé A Medina M. “Impact of the COVID-19 pandemic and related control measures on cancer diagnosis in Catalonia: a time-series analysis of primary care electronic health records covering about five million people.” In: *BMJ Open.* (2021). DOI: 10.1136/bmjopen-2020-047567.

[29] Palmer K Monaco A Kivipelto M Onder G Maggi S Michel JP Prieto R Sykara G Donde S. “The potential long-term impact of the COVID-19 outbreak on patients with non-communicable diseases in Europe: consequences for healthy ageing.” In: *Aging Clin Exp Res.* (Mar. 2020). DOI: 10.1007/s40520-020-01601-4.

[30] Rahmani A. Sabetkish N. “The overall impact of COVID-19 on healthcare during the pandemic: A multidisciplinary point of view.” In: (Oct. 2021). DOI: 10.1002/hsr2.386.

[31] Angelica Tiotiu et al. “Impact of the COVID-19 pandemic on the management of chronic noninfectious respiratory diseases”. In: *Expert Review of Respiratory Medicine* 15.8 (2021). PMID: 34253132, pp. 1035–1048. DOI: 10.1080/17476348.2021.1951707. eprint: https://doi.org/10.1080/17476348.2021.1951707. URL: https://doi.org/10.1080/17476348.2021.1951707.

[32] M. Azarpazhooh et al. “COVID-19 Pandemic and Burden of Non-Communicable Diseases: An Ecological Study on Data of 185 Countries”. In: *Journal of Stroke and Cerebrovascular Diseases* 29 (June 2020), p. 105089. DOI: 10.1016/j.jstrokecerebrovasdis.2020.105089.

[33] Palmer K Monaco A Kivipelto M Onder G Maggi S Michel JP Prieto R Sykara G Donde S. “The potential long-term impact of the COVID-19 outbrake on patients with non-communicable diseases in Europe: consequences for healthy ageing.” In: *Aging Clin Exp Res.* (June 2020), pp. 1189–1194. DOI: 10.1007/s40520-020-01601-4.

[34] Morros R. Bolíbar B Fina Avilés F. “Base de datos SIDIAP: La historia clínica informatizada de Atención Primaria como Fuente de información para La investigación epidemiológica.” In: (2019).

[35] Garcia-Gil M Elorza J-M Banque M et al. “Linking of primary care records to census data to study the association between socioeconomic status and cancer incidence in southern Europe: a nation-wide ecological study.” In: (Oct. 2014).

[36] Hyndman R.J. & Athanasopoulos G. *Forecasting: principles and practice, 2nd edition*. Melbourne, Australia: OTexts, 2018.

[37] Russell A. Poldrack. *Statistical Thinking for the 21st Century*. Self-published, 2018. URL: https://statsthinking21.org.

[38] World Medical Association. “World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects”. In: *JAMA* 310.20 (Nov. 2013), pp. 2191–2194. ISSN: 0098-7484. DOI: 10.1001/jama.2013.281053. eprint: https://jamanetwork.com/journals/jama/articlepdf/1760318/jsc130006.pdf. URL: https://doi.org/10.1001/jama.2013.281053.

[39] *DadesCovid*. https://dadescovid.cat/ap_diari?lang=eng. Accessed: 2022-05-22.

[40] *Momo*. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/MoMo/Paginas/MoMo.aspx. Accessed: 2022-05-22.

[41] María Ramos et al. “Relationship of diagnostic and therapeutic delay with survival in colorectal cancer: A review”. In: *European journal of cancer (Oxford, England : 1990)* 43 (Dec. 2007), pp. 2467–78. DOI: 10.1016/j.ejca.2007.08.023.

[42] Salvador Pita-Fernández et al. “Effect of diagnostic delay on survival in patients with colorectal cancer: A retrospective cohort study”. In: *BMC Cancer* 16 (Aug. 2016). DOI: 10.1186/s12885-016-2717-z.

[43] Schutte H. W. Heutink F. Wellenstein D. J. van den Broek G. B. van den Hoogen F. J. A. Marres H. A. M. van Herpen C. M. L. Kaanders J. H. A. M. Merkx T. M. A. W. “Impact of Time to Diagnosis and Treatment in Head and Neck Cancer: A Systematic Review.” In: *Otolaryngology–Head and Neck Surgery* (2020). DOI: https://doi.org/10.1177/0194599820906387.

[44] Alsallakh M.A. Sivakumaran S. Kennedy S. et al. “Impact of COVID-19 lockdown on the incidence and mortality of acute exacerbations of chronic obstructive pulmonary disease: national interrupted time series analyses for Scotland and Wales.” In: (2021). DOI: https://doi.org/10.1186/s12916-021-02000-w.

[45] Coma E Ferran M Méndez L Iglesias B Fina F Medina M. “Creation of a synthetic indicator of quality of care as a clinical management standard in primary care.” In: (Dec. 2013). DOI: 10.1186/2193-1801-2-51.

[46] *Recommendations from European Respiratory Society Group 9.1 (Respiratory function technologists/ Scientists) Lung function testing during COVID-19 pandemic and beyond*. https://static.physoc.org/app/uploads/2020/06/03140535/ERS-9.1-Statement-on-lung-function-during-COVID-19.pdf. Accessed: 2022-05-22.

[47] Klain A. Indolfi C. Dinardo G. et al. “Covid-19 and spirometry in this age.” In: (2022). DOI: https://doi.org/10.1186/s13052-022-01199-5.

[48] B. V. Murali2 Singla Rupak3 Koul Parvaiz Christopher Devasahayam Jesudas1 Mohan. “Pulmonary function testing during the COVID-19 pandemicin”. In: (Mar. 2021). DOI: 10.4103/lungindia.lungindia_738_20.

[49] Sofía Ruiz-Medina et al. “Significant Decrease in Annual Cancer Diagnoses in Spain during the COVID-19 Pandemic: A Real-Data Study”. In: *Cancers* 13.13 (June 2021), p. 3215. ISSN: 2072-6694. DOI: 10.3390/cancers13133215. URL: http://dx.doi.org/10.3390/cancers13133215.

[50] Talía Malagón Jean H. E. Yong Parker Tope Wilson H. Miller Jr. Eduardo L. Franco McGill. “Predicted long-term impact of COVID-19 pandemic-related care delays on cancer mortality in Canada.” In: *International Journal of Cancer.* (2021). DOI: https://doi.org/10.1002/ijc.33884.

[51] Hacker KA Briss PA Richardson L Wright J Petersen R. “COVID-19 and Chronic Disease: The Impact Now and in the Future.” In: (2021). DOI: http://dx.doi.org/10.5888/pcd18.210086.

[52] H.M. Peacock et al. “Decline and incomplete recovery in cancer diagnoses during the COVID-19 pandemic in Belgium: a year-long, population-level analysis”. In: *ESMO Open* 6.4 (2021), p. 100197. ISSN: 2059-7029. DOI: https://doi.org/10.1016/j.esmoop.2021.100197. URL: https://www.sciencedirect.com/science/article/pii/S2059702921001587.

[53] Jason K. Gurney Elinor Millar Alex Dunn Ruth Pirie Michelle Mako John Manderson et al. “The impact of the COVID-19 pandemic on cancer diagnosis and service access in New Zealand–a country pursuing COVID-19 elimination.” In: (2021). DOI: https://doi.org/10.1016/j.lanwpc.2021.100127.

Appendices

A APPENDIX I: ICD-10 codes

The codes considered for each of the studied diagnoses are presented below. Note that these codes belong to THE ICD-10-CM classification listed by the WHO.

Malignant neoplasms

C00.0, C00.1, C00.2, C00.4, C00.6, C00.8, C00.9, C01, C02.0, C02.1, C02.2, C02.3, C02.4, C02.8, C02.9, C03.0, C03.9, C04.1, C04.8, C04.9, C05.0, C05.1, C05.2, C05.8, C05.9, C06.0, C06.2, C06.80, C06.9, C07, C08.0, C08.9, C09.8, C09.9, C10.1, C10.3, C10.8, C10.9, C11.8, C11.9, C12, C13.8, C13.9, C14.0, C14.2, C14.8, C15.3, C15.4, C15.5, C15.9, C16.0, C16.1, C16.2, C16.3, C16.4, C16.5, C16.6, C16.8, C16.9, C17.0, C17.1, C17.3, C17.8, C17.9, C18.0, C18.1, C18.2, C18.3, C18.4, C18.5, C18.6, C18.7, C18.8, C18.9, C19, C20, C21.0, C21.8, C22.0, C22.1, C22.2, C22.3, C22.7, C22.9, C23, C24.0, C24.1, C24.8, C24.9, C25.0, C25.1, C25.2, C25.3, C25.4, C25.8, C25.9, C26.0, C26.9, C30.0, C30.1, C31.0, C31.3, C31.8, C32.0, C32.1, C32.2, C32.8, C32.9, C33, C34.00, C34.10, C34.2, C34.30, C34.80, C34.90, C37, C38.0, C38.1, C38.2, C38.3, C38.4, C38.8, C39.0, C39.9, C40.00, C40.10, C40.20, C40.30, C40.80, C40.90, C41.0, C41.1, C41.2, C41.3, C41.4, C41.9, C43.0, C43.10, C43.20, C43.30, C43.4, C43.59, C43.60, C43.70, C43.8, C43.9, C44, C44.00, C44.101, C44.201, C44.300, C44.40, C44.509, C44.601, C44.701, C44.80, C44.90, C44.91, C44.92, C44.99, C45.0, C45.1, C45.2, C45.7, C45.9, C46.0, C46.1, C46.2, C46.3, C46.7, C46.9, C47.0, C47.8, C47.9, C48.0, C48.2, C49.0, C49.10, C49.20, C49.3, C49.4, C49.5, C49.6, C49.8, C49.9, C50, C50.019, C50.212, C50.219, C50.312, C50.419, C50.619, C50.819, C50.919, C51.0, C51.1, C51.8, C51.9, C52, C53.0, C53.1, C53.8, C53.9, C54.0, C54.1, C54.2, C54.8, C54.9, C55, C56.9, C57.00, C57.4, C57.8, C57.9, C58, C60.0, C60.1, C60.8, C60.9, C61, C62.00, C62.10, C62.90, C63.7, C63.8, C63.9, C64.9, C65.9, C66.9, C67, C67.0, C67.1, C67.2, C67.3, C67.5, C67.6, C67.8, C67.9, C68.0, C68.8, C68.9, C69.00, C69.20, C69.30, C69.80, C69.90, C70.0, C70.9, C71.0, C71.1, C71.2, C71.3, C71.4, C71.5, C71.6, C71.7, C71.8, C71.9, C72.0, C72.30, C72.40, C72.9, C73, C76.0, C76.1, C76.2, C76.3, C76.40, C76.50, C76.8, C77.0, C77.1, C77.2, C77.3, C77.4, C77.8, C77.9, C78.00, C78.2, C78.4, C78.5, C78.6, C78.7, C78.80, C79.00, C79.10, C79.2, C79.31, C79.40, C79.51, C79.60, C79.70, C79.89, C79.9, C80.1, C81.10, C81.20, C81.30, C81.70, C81.90, C82.33, C82.90, C83.50, C83.70, C84.00, C84.10, C84.40, C85.10, C85.90, C88.0, C88.2, C88.3, C88.9, C90.00, C90.10, C90.20, C91.00, C91.10, C91.40, C91.90, C91.Z0, C92.00, C92.10, C92.30, C92.40, C92.50, C92.90, C92.Z0, C93.00, C93.10, C93.90, C94.80, C95.00, C95.10, C95.90, C96.2, C96.9, C96.A, C96.Z, D03.59, D06.0, D06.1, D06.7, D06.9, D07.4, D07.5, D07.60, D09.20, D09.8, D09.9, D45.

Diabetes

E11, E11.01, E11.21, E11.22, E11.29, E11.311, E11.319, E11.39, E11.40, E11.42, E11.43, E11.49, E11.51, E11.59, E11.610, E11.638, E11.649, E11.65, E11.69, E11.8, E11.9, E13.10, E13.29, E13.39, E13.49, E13.59, E13.641, E13.69, E13.8, E13.9.

Chronic Obstructive Pulmonary Disease (COPD)

J43.0, J43.1, J43.2, J43.8, J43.9, J44, J44.0, J44.1, J44.9.

Ischemic heart disease

I20, I20.0, I20.8, I20.9, I21, I21.0, I21.01, I21.02, I21.09, I21.1, I21.11, I21.19, I21.2, I21.29, I21.3, I21.4, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.8, I24.0, I24.1, I24.8, I24.9, I25, I25.10, I25.2, I25.41, I25.5, I25.6, I25.89, I25.9.

B APPENDIX II: Data preprocessing and flow charts

In this Appendix we are going to present the data preprocessing that has been performed for each disease. First of all, we are going to present how the categorization for the age variable has been carried out. Note that in this section just the results of the neoplasms are shown since the same procedure is used for each disease. After that, we present the flow charts which define the preprocessing applied to each disease. In other words, we show the number of diagnoses that we retrieved from the database and how we have excluded some as there was some missing information or some didn't meet the inclusion criteria.

Neoplasms' data categorization

Regarding age distribution, we explore its distribution taking into account all the cases in order to first delete the outliers, defined as people older than 120 years old, and then be able to categorise it. As exposed in methods, the 25th, 50th and 75th quartiles are used as categorisation thresholds.

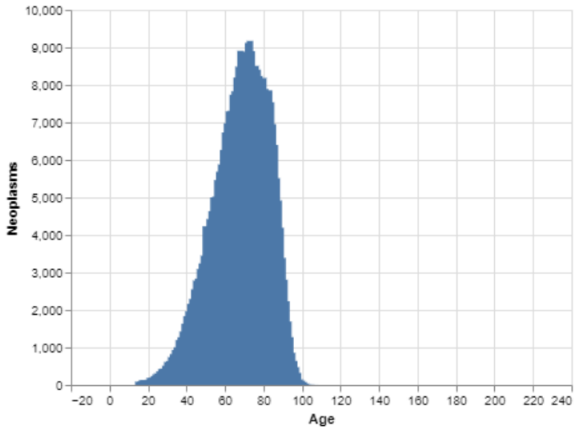


Figure 19: Number of neoplasms according to age at diagnosis.

As shown in Figure 19, we can see that most neoplasms are found on people around 70 years old. From this Figure we can observe that the distribution has a Gaussian form with a maximum at the age of 72 and a longer tail on the left.

Empirically, the mean of the people diagnosed with neoplasms is 67.7 years old, the median is 69.0 and the 25 and 75 percentiles are 58 and 79 respectively. All the mentioned values can be observed in the boxplot 20.

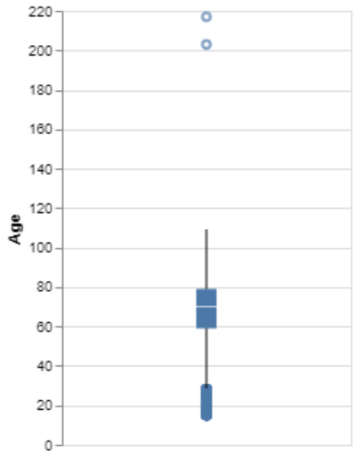


Figure 20: Distribution of age at which people got diagnosed with a neoplasm.

B.1 Flow charts

As mentioned before, here cleaning processes for each disease are depicted.

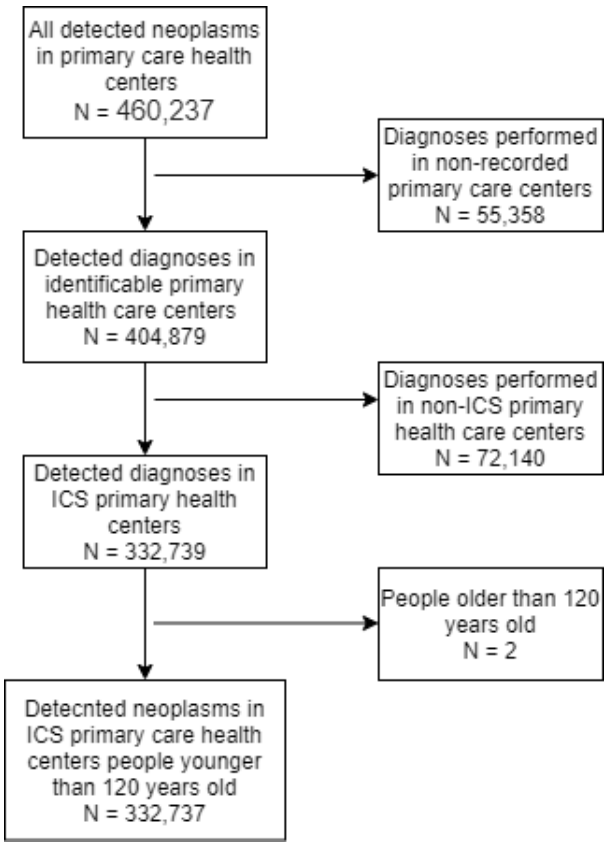


Figure 21: Neoplasms' data cleaning pipeline.

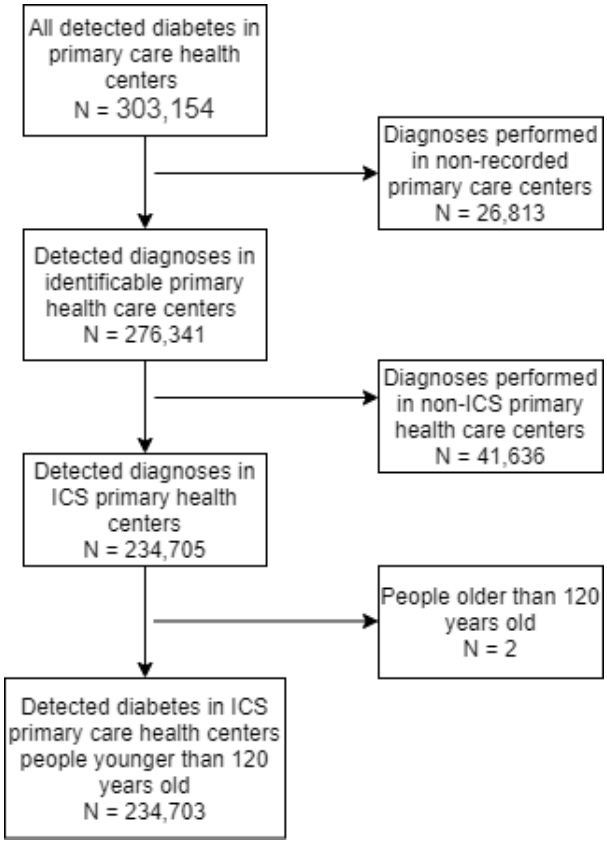


Figure 22: Diabetes' data cleaning pipeline.

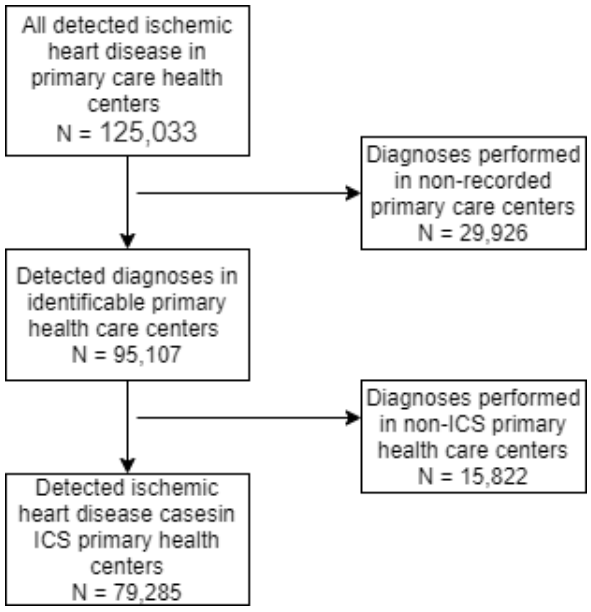


Figure 23: Ischemic heart disease's data cleaning pipeline.

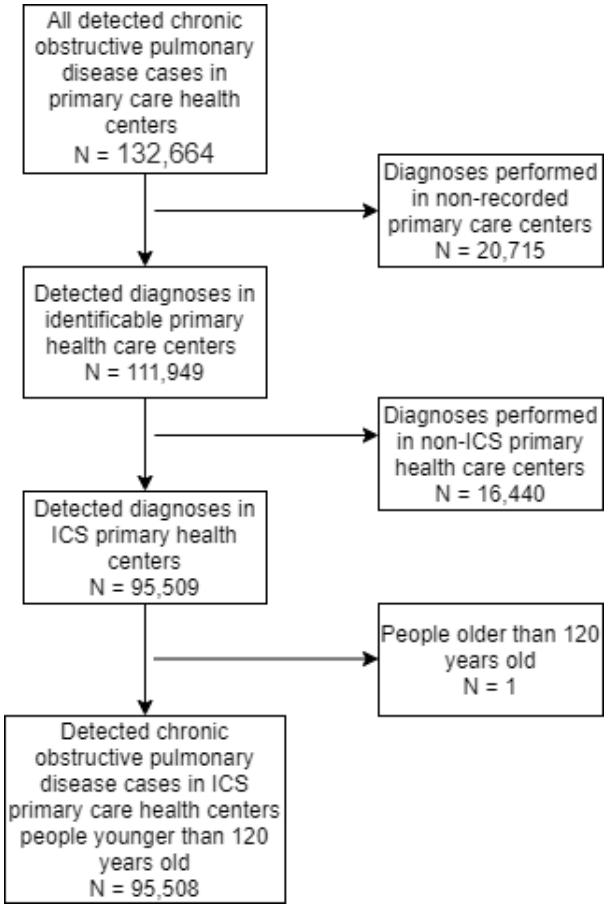


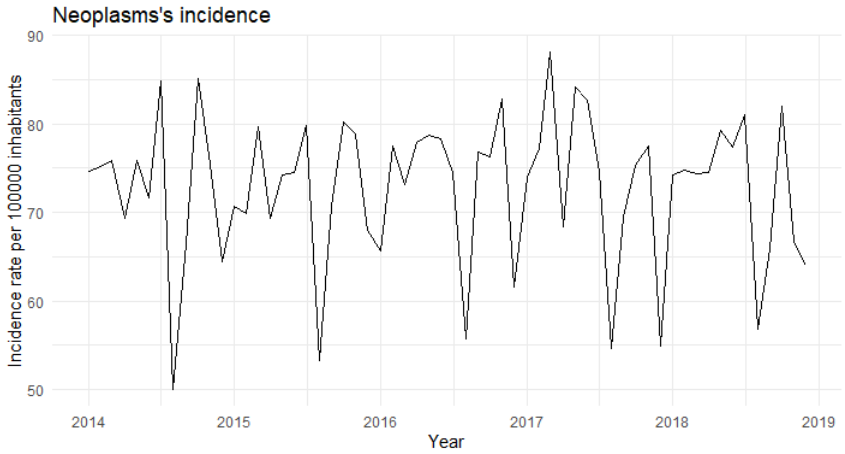
Figure 24: Chronic obstructive pulmonary disease's data cleaning pipeline.

C APPENDIX III: Details and validation of the models for each disease

C.1 Neoplasms

C.1.1 Adjusted model specifications

Our global neoplasm time series, during the study period, has the following pattern:



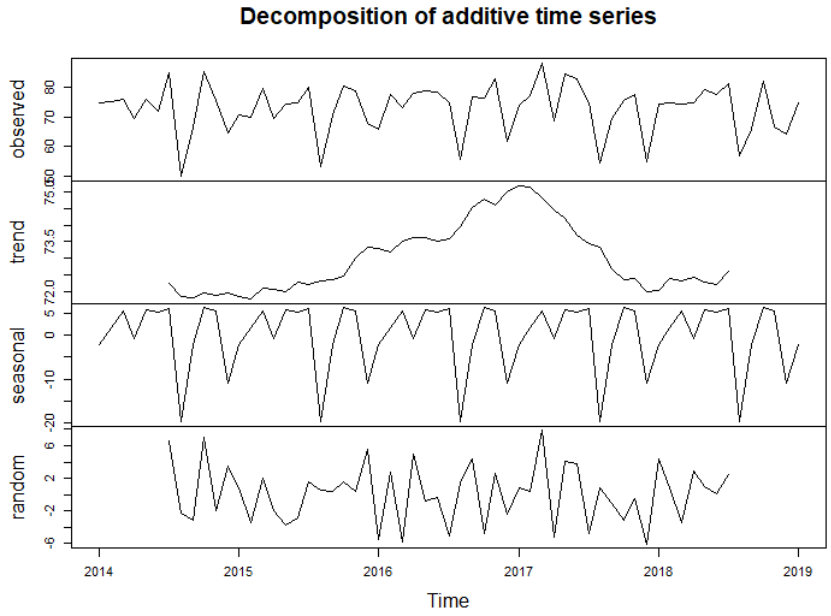
Using a time series regression and the following formulas, we were able to calculate the predicted incidence for the research period. $Y_t = 71.56 + 0.01t + 3.03s_{t2} + 6.35s_{t3} + 0.02s_{t4} + 6.53s_{t5} + 5.01s_{t6} + 7.03s_{t7} - 17.83s_{t8} - 2.05s_{t9} + 7.84s_{t10} + 4.29s_{t11} - 9.37s_{t12} + e_t$ Where t is the instant time, s_{ti} is the seasonal period and e_t is the random error.

The following is a summary of the regression model's estimated results:

```
Call:
tslm(formula = dx_ts ~ trend + season)

Residuals:
   Min     1Q Median     3Q    Max
-9.787 -3.194  0.436  2.159  9.751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.56484    2.14197  33.411  < 2e-16 ***
trend         0.01127    0.03353   0.336  0.73820
season2       3.03096    2.78781   1.087  0.28249
season3       6.34827    2.78842   2.277  0.02740 *
season4       0.02165    2.78943   0.008  0.99384
season5       6.53475    2.79084   2.342  0.02350 *
season6       5.01109    2.79265   1.794  0.07918 .
season7       7.03238    2.79486   2.516  0.01534 *
season8     -17.82793    2.79748  -6.373 7.32e-08 ***
season9      -2.05335    2.80049  -0.733  0.46707
season10      7.84216    2.80390   2.797  0.00745 **
season11      4.28967    2.80770   1.528  0.13326
season12     -9.37219    2.81191  -3.333  0.00168 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.408 on 47 degrees of freedom
Multiple R-squared:  0.7794,    Adjusted R-squared:  0.7231
F-statistic: 13.84 on 12 and 47 DF,  p-value: 1.234e-11
```

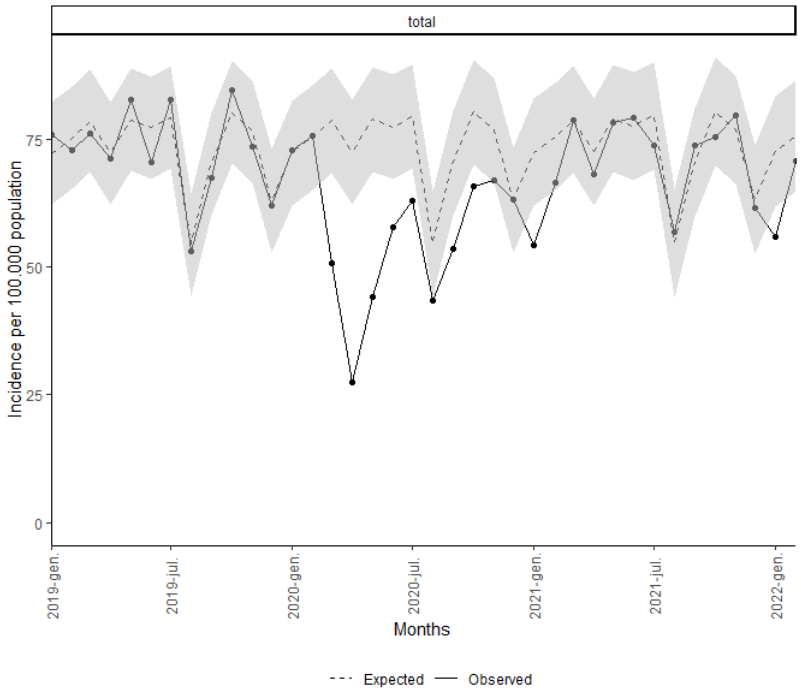
From the summary, we can observe that the coefficient associated to the trend is not statistically significant. In addition, there are some seasons that are not statistically different from the baseline. However, we opted to maintain them in our model because the time series decomposition shows a trend component as well as seasonality. These

facts can be seen in the graph below, as well as in the summary, because the seasonality is statistically significant for season 8 or 12, for example, vacation periods during August and December.



C.1.2 Model validation

In order to validate the model we need to make sure that the model fits the validation year correctly and also that the assumptions made by the model are met. Regarding the validation period, in the following plot we can see that the model fits properly the validation period as all the observed points during it fall inside the IC95%.

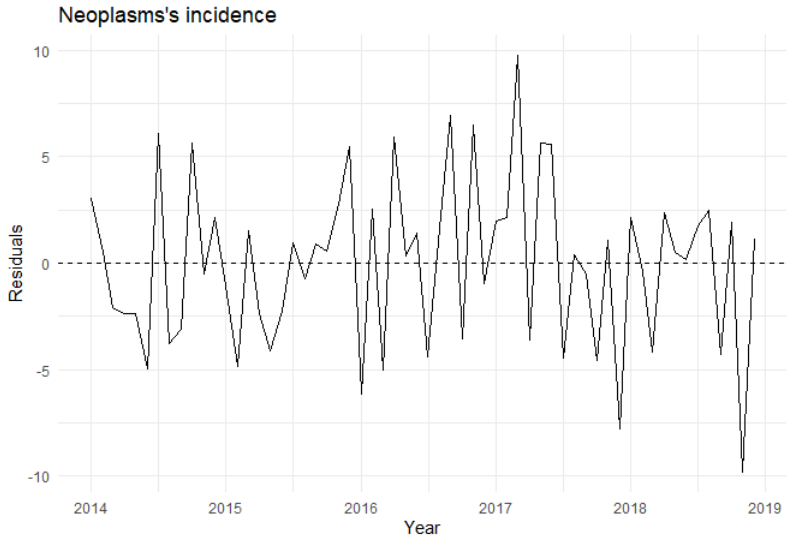


Our time series linear regression model makes the following assumptions:

1. The residuals are uncorrelated.
2. The residuals have zero mean.
3. The residuals have constant variance.
4. The residuals are normally distributed.

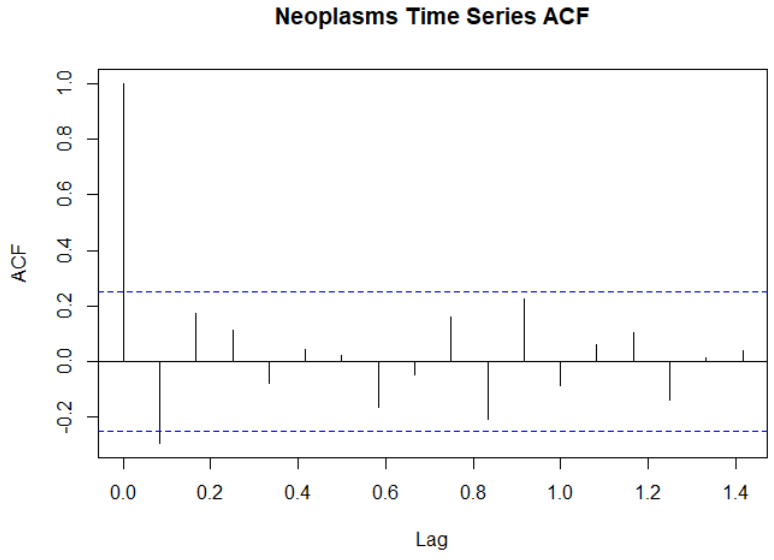
These assumptions can be validated with the following figures and tests:

Zero mean and constant variance residuals



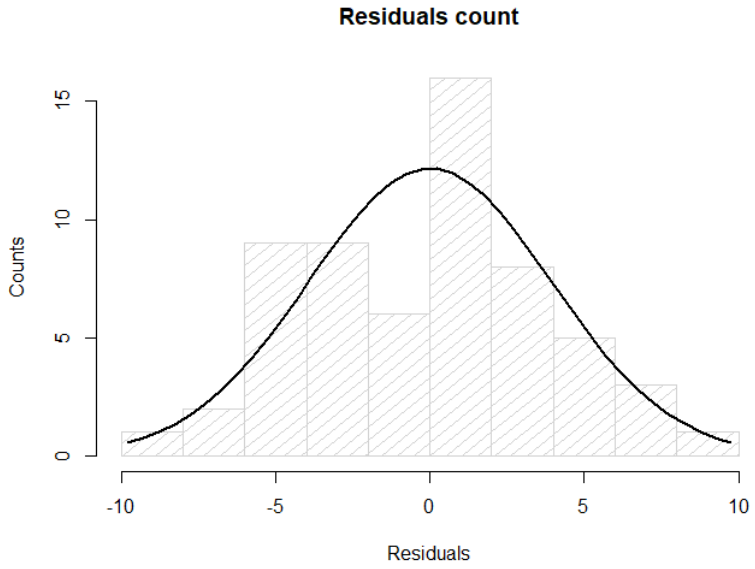
We can see that the residuals are around 0 in this time plot of the residuals for the 5 years included in the study period. We may also see that variance remains constant throughout time. As a result of this figure, we can conclude that our model meets the b and c assumptions.

Uncorrelated residuals



The Neoplasms Time Series ACF figure shows autocorrelation of residuals. These residuals have to be uncorrelated and, as we can observe, they are not significant (except for lag 2), hence uncorrelated. In order to validate the non-autocorrelation of the residuals, Ljung-Box and Box-Pierce tests have been calculated and non-significant p-values level, 0.1077 and 0.2925 respectively with 20 degrees of freedom, have been observed. Hence, our residuals can not be differentiated from white noise.

Normally distributed residuals

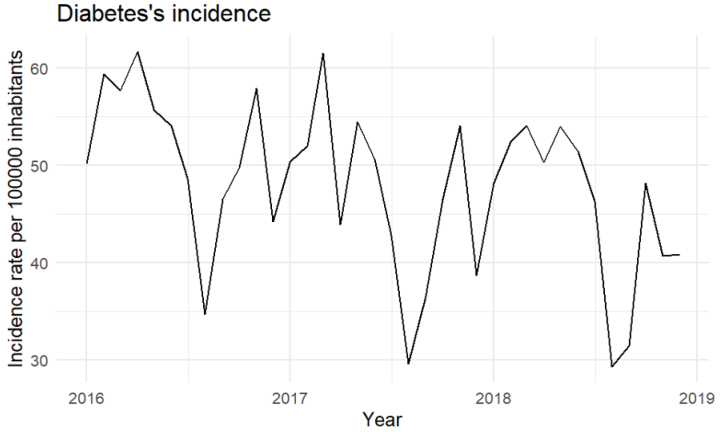


Finally, the histogram of residuals displays a near-Normal distribution that is 0-centred. We may thus assume that the residuals in our model have a normal distribution with a mean of 0 and a constant variance.

C.2 Diabetes

C.2.1 Adjusted model specifications

Our global diabetes time series, during the study period, has the following pattern:

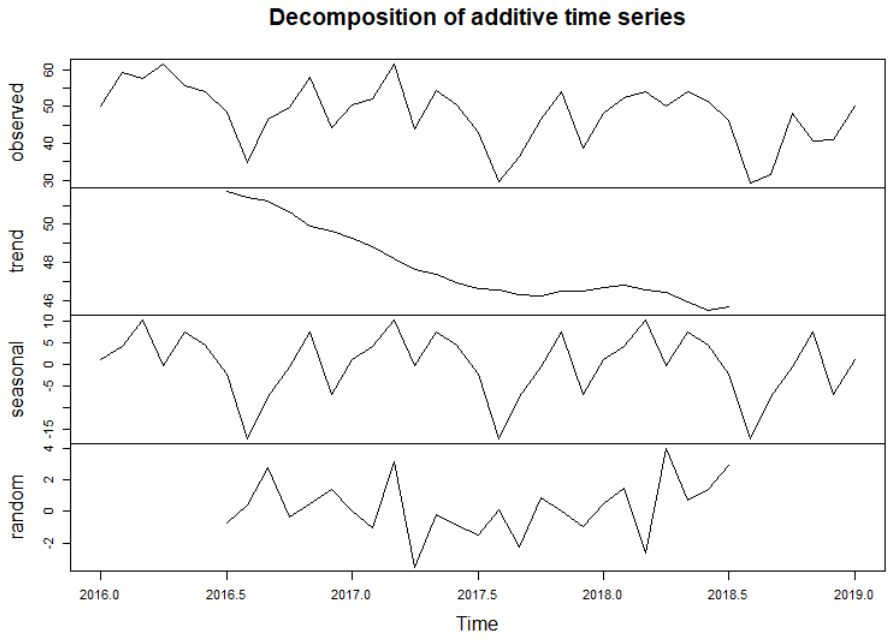


Using a time series regression and the following formulas, we were able to calculate the predicted incidence for the research period. $Y_t = 49.55 + 5.05s_{t2} + 8.20s_{t3} + 2.40s_{t4} + 5.14s_{t5} + 2.43s_{t6} - 3.64s_{t7} - 18.31s_{t8} - 11.39s_{t9} - 1.42s_{t10} + 1.33s_{t11} - 8.29s_{t12} + e_t$ Where t is the instant time, s_{ti} is the seasonal period and e_t is the random error.

The following is a summary of the regression model's estimated results:

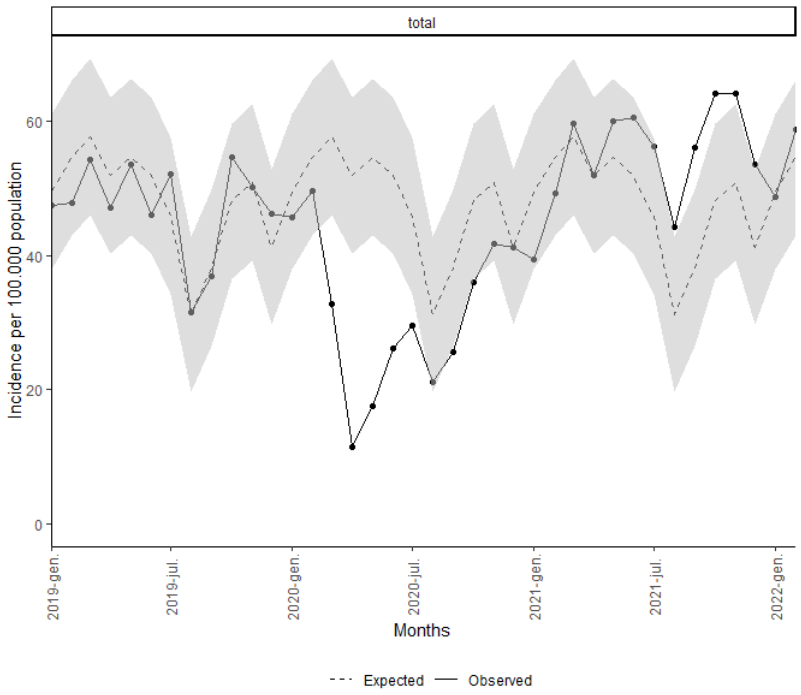
```
Call:
tslm(formula = dx_ts ~ season)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1776  -1.7459  -0.3283   2.2020   9.7127

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.550      2.815  17.604 3.17e-15 ***
season2        5.053      3.981   1.269 0.216491
season3        8.195      3.981   2.059 0.050529 .
season4        2.403      3.981   0.604 0.551744
season5        5.138      3.981   1.291 0.209093
season6        2.433      3.981   0.611 0.546826
season7       -3.654      3.981  -0.918 0.367833
season8      -18.310      3.981  -4.600 0.000115 ***
season9      -11.394      3.981  -2.862 0.008588 **
season10      -1.415      3.981  -0.356 0.725289
season11       1.331      3.981   0.334 0.741070
season12      -8.287      3.981  -2.082 0.048193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.875 on 24 degrees of freedom
Multiple R-squared:  0.7745,    Adjusted R-squared:  0.6712
F-statistic: 7.494 on 11 and 24 DF,  p-value: 2.072e-05
```

From the summary, we can observe that seasonalities related to seasons 3, 8 and 9 and 12. However, there are some seasons that are not statistically different from the baseline. Nonetheless, we opted to maintain them in our model because the time series decomposition shows a clear seasonality. These facts can be seen in the graph below, as well as in the summary, because the seasonality is statistically significant for season 8 or 12, for example, holiday periods during August and December.



C.2.2 Model validation

In order to validate the model we need to make sure that the model fits the validation year correctly and also that the assumptions made by the model are met. Regarding the validation period, in the following plot we can see that the model fits properly the validation period since all the observed monthly incidence rates are in the expected range.

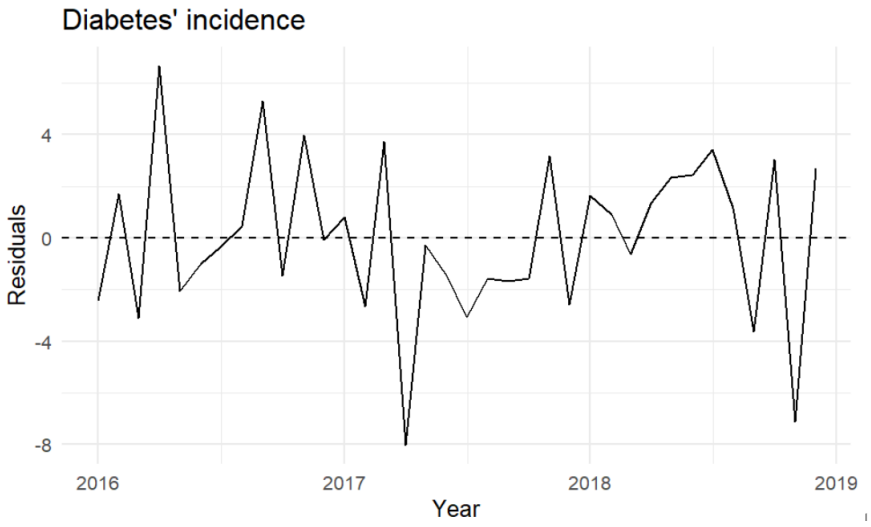


Our time series linear regression model makes the following assumptions:

1. The residuals are uncorrelated.
2. The residuals have zero mean.
3. The residuals have constant variance.
4. The residuals are normally distributed.

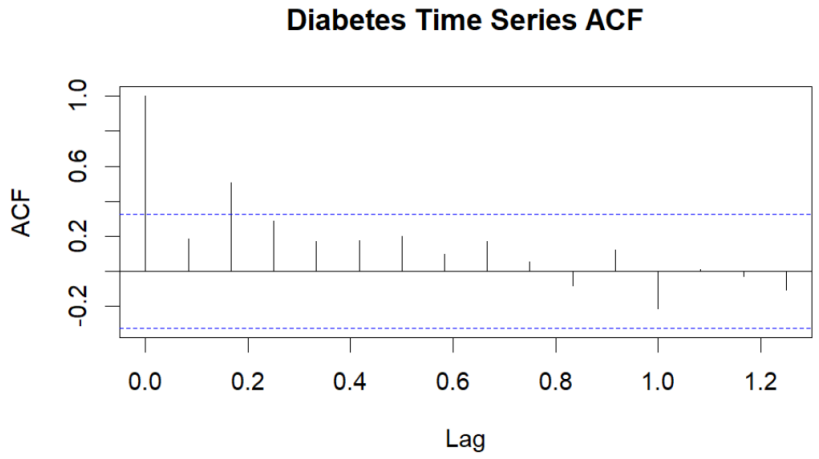
These assumptions can be validated with the following figures and tests:

Zero mean and constant variance residuals



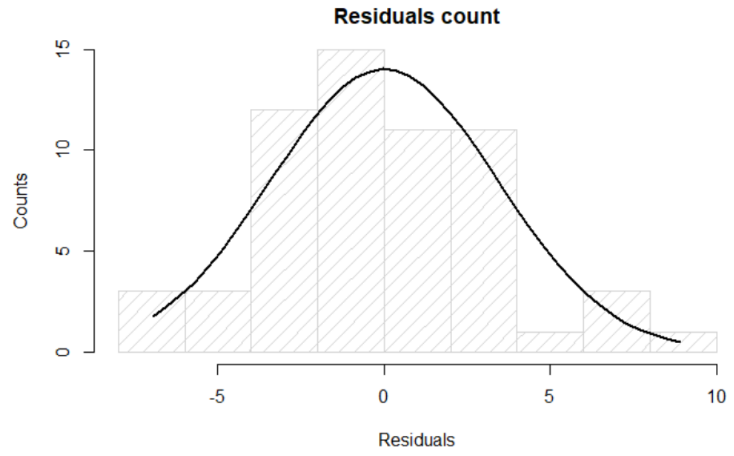
We can see that the residuals are around 0 in this time plot of the residuals for the 3 years included in the study period. We may also see that variance remains constant throughout time. As a result of this figure, we can conclude that our model meets the 2 and 3 assumptions.

Uncorrelated residuals



The Diabetes' Time Series ACF figure shows autocorrelation of residuals. These residuals have to be uncorrelated and, as we can observe, they are not significant (except for lag 3), hence uncorrelated. In order to validate the non-autocorrelation of the residuals, Ljung-Box and Box-Pierce tests have been calculated and non-significant p-values levels, 0.0669 and 0.2304 respectively with 20 degrees of freedom, have been observed. Hence, our residuals can not be differentiated from white noise.

Normally distributed residuals

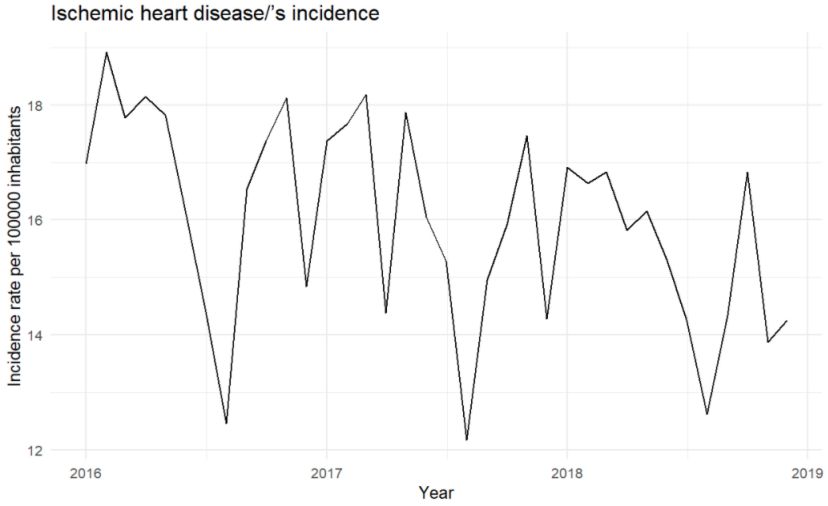


Finally, the histogram of residuals displays a near-Normal distribution that is 0-centred. We may thus assume that the residuals in our model have a normal distribution with a mean of 0 and a constant variance.

C.3 Ischemic heart disease

C.3.1 Adjusted model specifications

Our global ischemic heart disease time series, during the study period, has the following pattern:

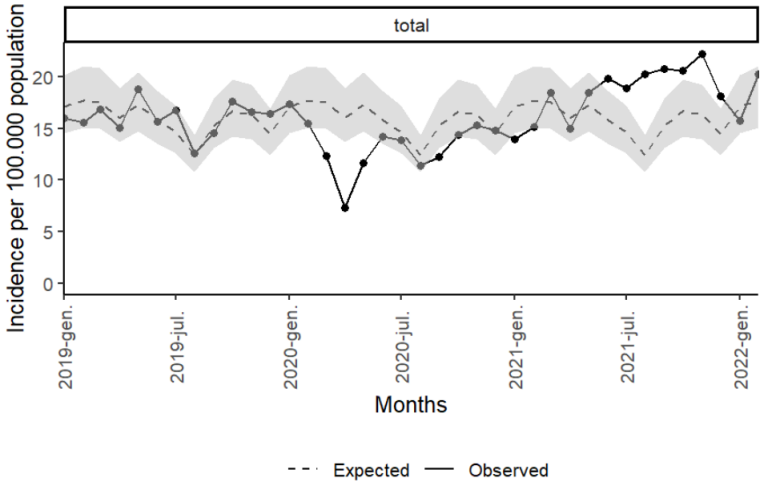


Using a time series regression and the following formulas, we were able to calculate the predicted incidence for the research period.

$$Log(log(Y_t)) = 1.059 - 0.001s_{t1} + 0.014s_{t2} + 0.012s_{t3}$$
$$-0.020s_{t4} + 0.008s_{t5} - 0.022s_{t6} - 0.049s_{t7} - 0.111s_{t8}$$
$$-0.032s_{t9} + 0.003s_{t10} - 0.004s_{t11} - 0.048s_{t12} + e_t$$

Where t is the instant time, s_{ti} is the seasonal period and e_t is the random error. Note that in this case a double differenciaciation has been used in order to model the time

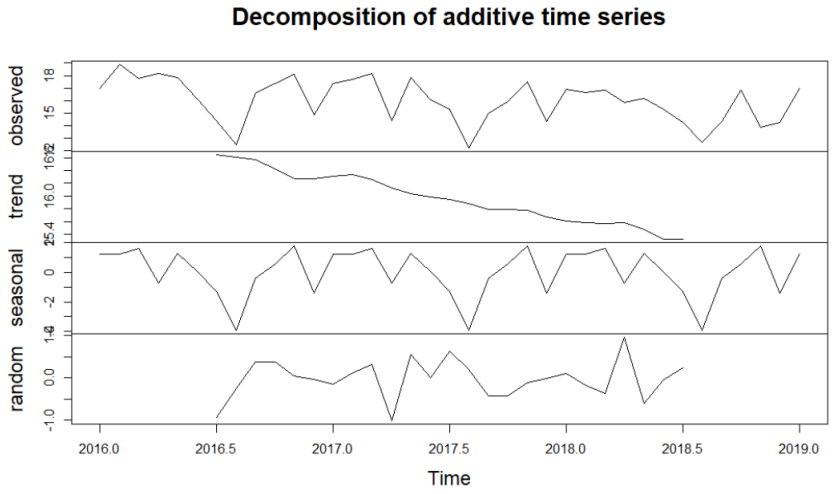
series and that by using it we can validate the time series in the training period as shown in the following figure:



The following is a summary of the regression model's estimated results:

```
Call:
tslm(formula = dx_ts ~ trend + season)

Residuals:
      Min        1Q    Median        3Q       Max
-0.046126 -0.008056 -0.000453  0.010905  0.029890

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0585844  0.0124192  85.238  < 2e-16 ***
trend       -0.0011826  0.0003417  -3.461  0.00212 **
season2      0.0137815  0.0164050   0.840  0.40951
season3      0.0124333  0.0164157   0.757  0.45650
season4     -0.0195034  0.0164334  -1.187  0.24742
season5      0.0082617  0.0164583   0.502  0.62046
season6     -0.0215827  0.0164902  -1.309  0.20352
season7     -0.0489446  0.0165291  -2.961  0.00700 **
season8     -0.1111093  0.0165749  -6.703 7.73e-07 ***
season9     -0.0316822  0.0166277  -1.905  0.06931 .
season10     0.0025522  0.0166872   0.153  0.87978
season11    -0.0041871  0.0167536  -0.250  0.80487
season12    -0.0477106  0.0168266  -2.835  0.00937 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary, we can observe that seasonalities and trends are statistically significant. In addition, these patterns are shown in the decomposition figure.



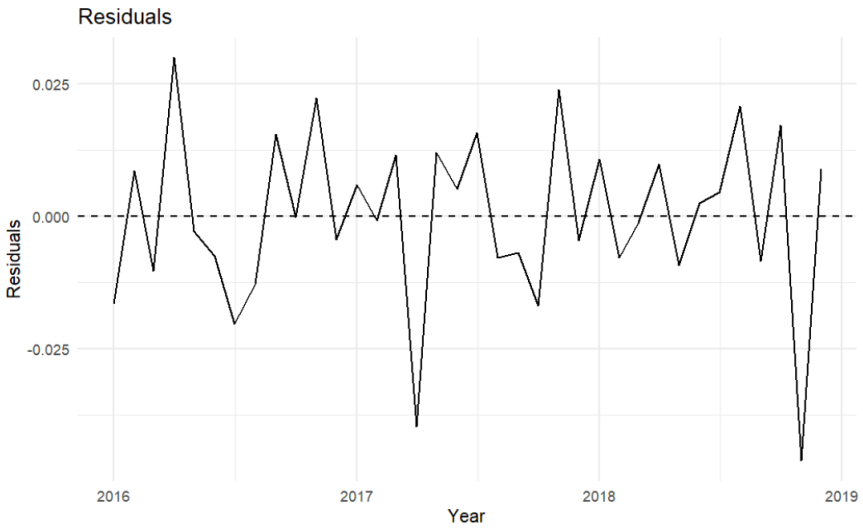
C.3.2 Model validation

Our time series linear regression model makes the following assumptions:

1. The residuals are uncorrelated.
2. The residuals have zero mean.
3. The residuals have constant variance.
4. The residuals are normally distributed.

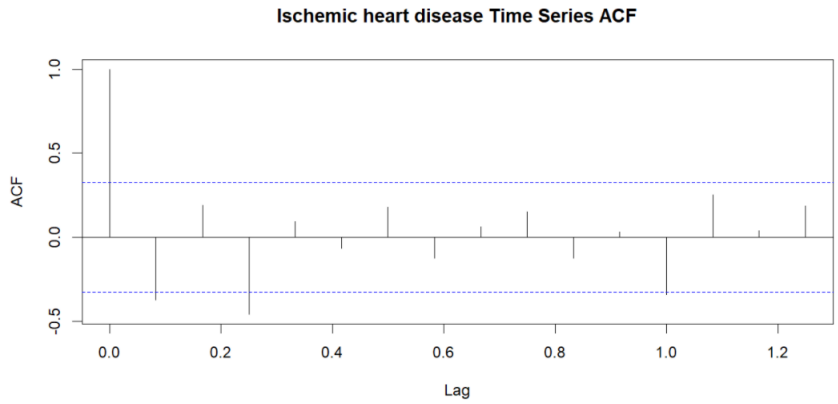
These assumptions can be validated with the following figures and tests:

Zero mean and constant variance residuals



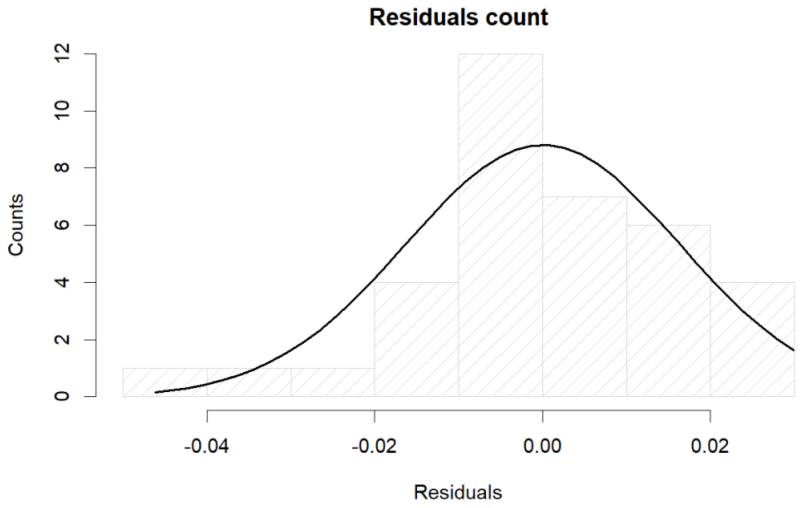
We can see that the residuals are around 0 in this time plot of the residuals for the 3 years included in the study period. We may also see that variance remains constant throughout time. As a result of this figure, we can conclude that our model meets the b and c assumptions.

Uncorrelated residuals



The Ischemic heart disease' Time Series ACF figure shows autocorrelation of residuals. These residuals have to be uncorrelated and, as we can observe, they are significant, hence correlated. In order to validate the non-autocorrelation of the residuals, Ljung-Box and Box-Pierce tests have been calculated and significant p-values levels, 1.93e-5 and 0.013 respectively with 20 degrees of freedom have been observed. Hence, our residuals can be differentiated from white noise.

Normally distributed residuals

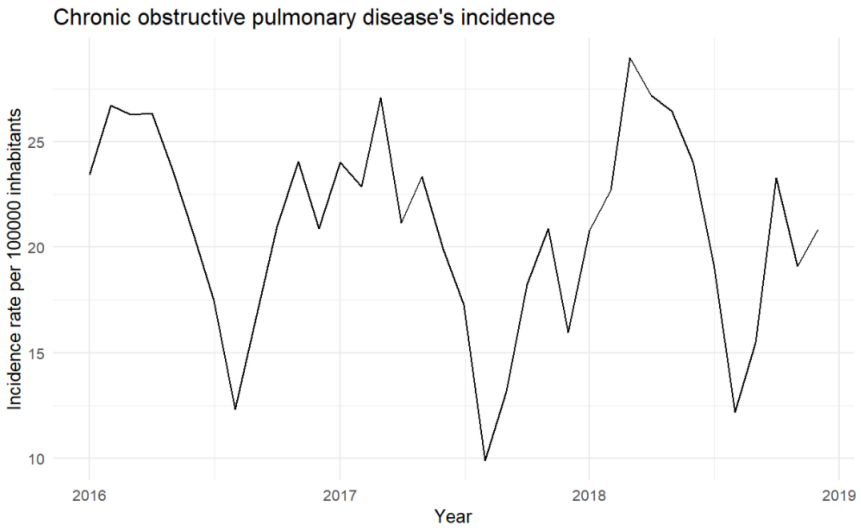


Finally, the histogram of residuals does not display a near-Normal Distribution. It shows a distribution that is close to 0-centred. However, we can see that the residuals are bigger on the right than on the left. Therefore, no Normal Distribution is observed. We may thus conclude that we cannot validate the model since no normally distributed residuals nor non-correlated residuals are observed. Hence, we opted to study the incidence rate using another suitable method, in this case mean difference with the Wilch test.

C.4 Chronic obstructive pulmonary disease

C.4.1 Adjusted model specifications

Our global chronic obstructive pulmonary disease time series, during the study period, has the following pattern:



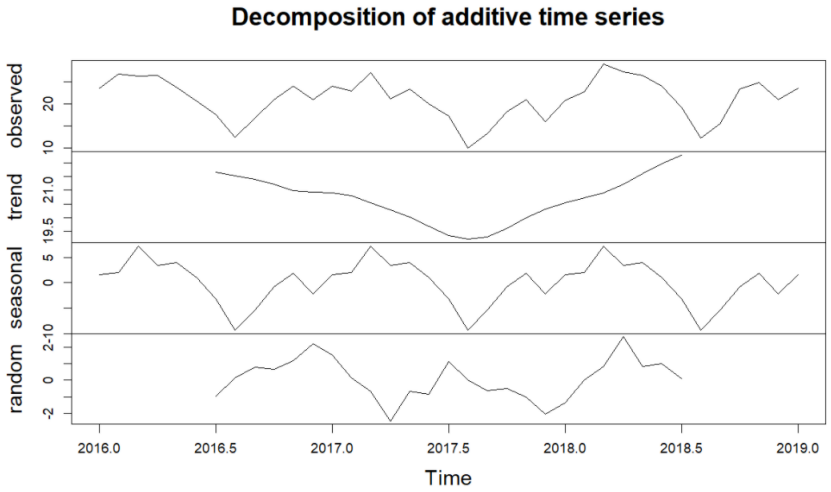
Using a time series regression, the tsclean function of R to replace outliers, with linear interpolation, and the following formulas, we were able to calculate the predicted incidence for the research period.

$$Y_t = 22.72 + 0.0026t + 1.35s_{t2} + 4.70s_{t3} + 2.15s_{t4} + 1.72s_{t5}$$
$$-1.28s_{t6} - 4.78s_{t7} - 11.31s_{t8} - 7.64s_{t9} - 1.94s_{t10} - 1.43s_{t11} - 3.53s_{t12} + e_t$$

Where t is the instant time, s_{ti} is the seasonal period and e_t is the random error. The following is a summary of the regression model's estimated results:

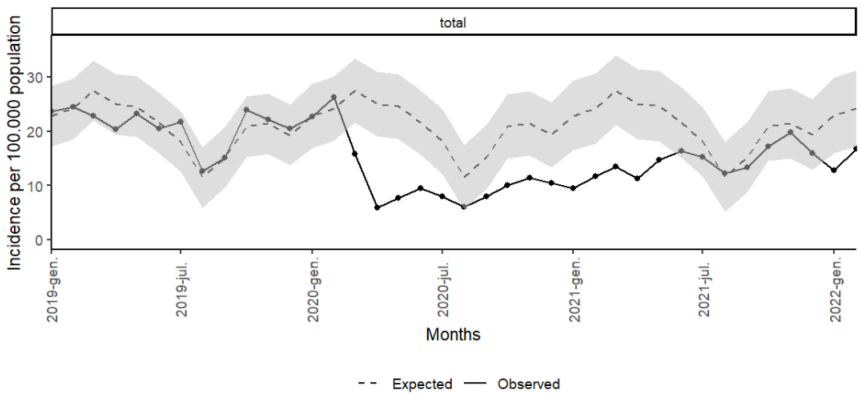
```
Call:
tslm(formula = dx_ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7532 -1.2689 -0.0789  1.5182  2.7403

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.724539   1.358943  16.722  2.3e-14 ***
trend         0.002566   0.037390   0.069 0.945877
season2       1.352877   1.795086   0.754 0.458702
season3       4.697758   1.796254   2.615 0.015470 *
season4       2.154357   1.798199   1.198 0.243093
season5       1.720548   1.800917   0.955 0.349324
season6      -1.281502   1.804407  -0.710 0.484715
season7      -4.775663   1.808663  -2.640 0.014621 *
season8     -11.314509   1.813681  -6.238  2.3e-06 ***
season9      -7.644959   1.819452  -4.202 0.000341 ***
season10     -1.942538   1.825972  -1.064 0.298444
season11     -1.429888   1.833230  -0.780 0.443354
season12     -3.534851   1.841220  -1.920 0.067365 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.198 on 23 degrees of freedom
Multiple R-squared:  0.8558,    Adjusted R-squared:  0.7806
F-statistic: 11.38 on 12 and 23 DF,  p-value: 5.443e-07
```

From the summary, we can observe that the trend is statistically non-significant. Therefore, we have decided to remove it from our model. Regarding seasonality, we can observe that although there are some seasons that are not statistically different from the baseline, we opted to maintain them in our model because the time series decomposition shows a clear seasonality. These facts can be seen in the graph below.



C.4.2 Model validation

In order to validate the model we need to make sure that the model fits the validation year correctly and also that the assumptions made by the model are met. Regarding the validation period, in the following plot we can see that the model fits properly the validation period since the observed incidence rates belong to the expected range.

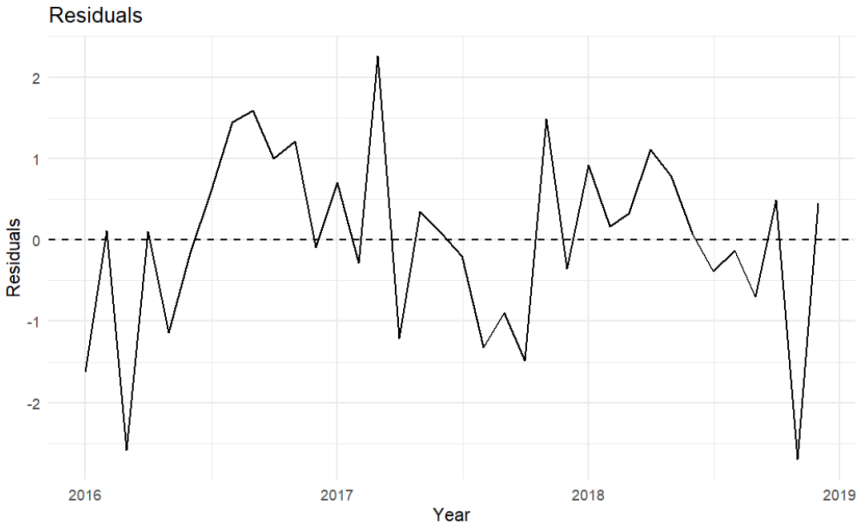


Our time series linear regression model makes the following assumptions:

1. The residuals are uncorrelated.
2. The residuals have zero mean.
3. The residuals have constant variance.
4. The residuals are normally distributed.

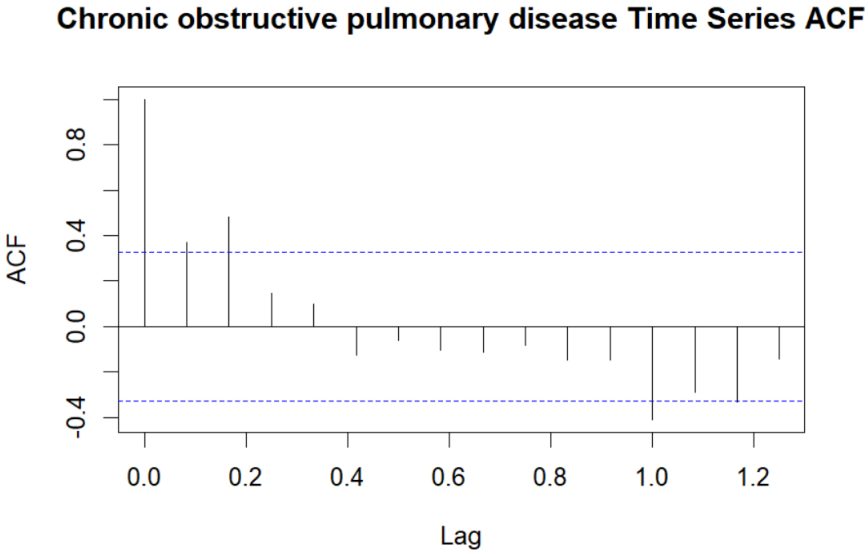
These assumptions can be validated with the following figures and tests:

Zero mean and constant variance residuals



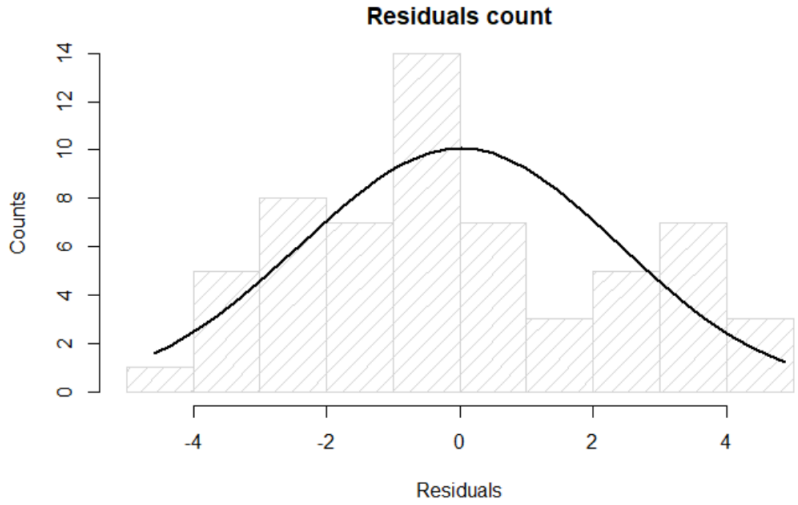
We can see that the residuals are around zero in this time plot of the residuals for the 3 years included in the study period. We may also see that variance remains constant throughout time but at the beginning of 2018. As a result of this figure, we can conclude that our model meets the b and c assumptions.

Uncorrelated residuals



The chronic obstructive pulmonary disease Time Series ACF figure shows autocorrelation of residuals in lags 2, 3 and 13. These residuals have to be uncorrelated and, as we can observe, apart from the lags commented before they are not significant, hence uncorrelated. In order to validate the non-autocorrelation of the residuals, Box-Pierce test has been calculated and a significant p-value level 0.076 with 20 degrees of freedom, has been observed. Hence, our residuals can not be differentiated from white noise.

Normally distributed residuals



Finally, the histogram of residuals displays a near-Normal distribution that is 0-centred with a passing tail in the right. We may thus assume that the residuals in our model have a normal distribution with a mean of 0 and a constant variance.

D APPENDIX IV: Number of spirometries

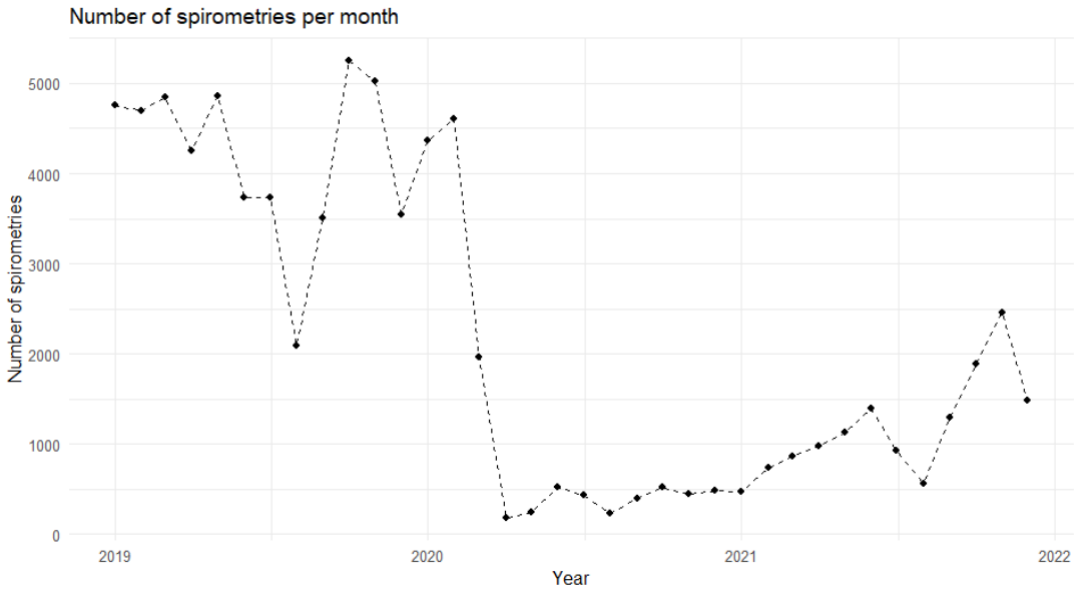


Figure 25: Number of spirometries per month.