# Analysis and Visualization of Incidents of Communicable Diseases

Nuno Ricardo Gomes Pires
nuno.gomes.pires@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

July 2017

**Abstract**

The Portuguese Directorate-General for Health (DGS) has an information system for reporting cases of mandatory notification surveillance, SINAVE, the National System of Epidemiological Surveillance. SINAVE is used for monitoring epidemiological data, but does provide any kind of visualization for that data. This dissertation presents a complementary system, the eVD Lab (E- communicable Diseases Surveillance), which provides an immediate overview of the incidence of communicable diseases. The eVD Lab enables visualization of information on several variables of analysis, such as disease, location, age and gender, using the potentialities of several visual elements, such as heatmap, line chart and choropleth. An evaluation of the system demonstrated good usability and good performance in the proposed tasks, and its usefulness for analysis in the DGS.
**Keywords:** Epidemiological surveillance; Visualization information; Real-time surveillance

## 1. Introduction

In a globalized world where are more and more trips and the transmission of infectious diseases is a major concern, there is a need to control and monitor the spread of disease. In this regard, there has been an increase in the development of computer applications to monitor data on various communicable diseases. Worldwide, there are a number of entities in the short term that aim to improve disease control systems, such as CDC (Center for Disease Control and Prevention of United States) and ECDC (European Center for Disease Control and Prevention) and who want to be at the forefront of disease control and prevention. To this end, they want to modernize data collection and visualization systems so that information on the state of public health is of a better quality and allows rapid and incisive responses to contain epidemiological diseases.

CDC and ECDC have systems that show some transmissible disease data and although they are not yet robust and complex applications, are a first step in achieving the goal of having a system capable of providing greater coverage in the control and surveillance of communicable diseases. Like CDC and ECDC, DGS in Portugal also aims to improve the control of epidemiological diseases. For this DGS implemented a system, SINAVE. In these systems are found personal data from patients with symptoms of diseases of mandatory surveillance, namely date of symptoms, illness, date of birth and gender. These data allow to make a study about the overview of the incidence of notifiable diseases.

SINAVE is a surveillance system in public health, which identifies high-risk situations, collects, updates, analyzes and disseminates data on communicable diseases. Since 1 January 2015 it is compulsory to use to notify communicable diseases. This ensures the database that supports the system, being possible to collect data from various regions of the country. The SINAVE consists of two parts the clinical SINAVE (SINAVEmed) and the Laboratory SINAVE (SINAVElab). At SINAVE in clinical data relating to visits and observations hospitals, being added to the system by doctors when diagnosing a notifiable disease. In SINAVElab the data relate to the results of laboratory analyzes that are done to confirm the existence of communicable diseases for which declaration is mandatory. The work is focused on the data from SINAVElab, being that simultaneously, the data from SINAVEmed were the focus of another master thesis by Daniela Pimentel de Oliveira, a student in the Msc in Biomedical Engineering - Branch of Clinical Engineering of the Universidade do Minho, having as objective the development of a similar system just for this information. In SINAVElab listed data such as the date of clinical analysis, disease, date of birth and gender of the wearer, as well as place of residence. These data are analyzed in order to obtain information about the incidence of notifiable diseases at a national level. However, to have a better perception and interpretation of the data, it is necessary to create the means that allow your viewing.

With the creation of information systems that al-

low the visualization through charts, maps, tables, and other visual elements, the task of detecting patterns, peaks and epidemics is facilitated, as can be seen in a simple and quick some abnormality, such as an increase in the number of cases of a particular disease or a seasonal peak through a line graph or even which regions most affected by a choropleth. To ensure that these systems can be used and contain relevant information must be supported by reliable data. The data on which these systems are based are data collected in hospital and clinical centers.

### 1.1. Goals

So that the work has quality and is well defined, were outlined some goals. The main goal established was the development of a system of analysis and visualization of incidence of communicable diseases.

For this reason, the system should allow::

- View information about the incidence of communicable diseases by regions, age groups and gender.
- Show the incidence of notifications, helping in the identification of extreme values for the control and prevention of outbreaks.
- View trends and developments over time.
- The export of data for analysis in the DGS and for external analysis.
- The public view of information on communicable diseases in Portugal.

## 2. Background

In this section are analyzed some systems of epidemiological data. VoroGraph is a tool for epidemiological analysis that allows to analyze the incidence and spread of disease in relation to population density and other demographic conditions in geographic scales ranging from international flights the local movements [1]. To represent this information, were considered several techniques, including visual maps, CVT and meta-based layout in CVT. The visual map uses classifications at the borders between neighboring regions to show local relations. The CVT is a specific case of Voronoi Diagrams. The Voronoi diagram is the partitioning of a plan in various regions based on the distance to a specific point of the plan. In the case of the CVT, the point of each region is the average of points of the original plan. The CVT allows transforming the map in the form of filling in the space, preserving relative positions, so as to highlight properties of the region as well as the local relations. Finally, the meta-layout is based on the CVT showing the aggregate of long distance beyond the local relations.

In terms of design uses a Border-Encoded Map which, through the clear identification of boundaries on a map, allows the visualization of the movements between regions (referred to as "basin") are contiguous, using sizes and colors to represent the ratio of population infected. The size of the border encodes the total number of movements between adjacent basins while the color, on a scale from white to red shows how many people have been infected. It also uses Morphing CVT, i.e., turns the map into CVT to cope with the limitations of codifications of border. Optimized CVT have several properties that make the information effective in viewing tools. Still using Labeling, Border Encodings, Cvt Meta-Layout, animated transitions and timeline.

The effective detection and response to outbreaks of infectious diseases depends upon the ability to capture and analyze information and how employees of public bodies are able to respond to this information. The Epinome, developed by [2], is an integrated system of visual research and analytics. This system has a dynamic environment that involves perfectly and fits the tasks of users and their needs. It features four paradigms of user-interaction in public health: visual display of evolution, perfect integration between different views, multiple views freely coordinate and direct interaction with the data.

Another example of these systems is the GIDEON. In GIDEON is possible to find some capabilities that an information system for infectious diseases must possess in order to be useful for the end user [3]. First, it is necessary that the system is comprehensive in order to encompass the clinical and epidemiological characteristics of all infectious diseases, all human pathogenic and like all vaccines and anti-infective. The program should be flexible and should be updated in "real time", besides allowing the modification of data by the development team/maintenance how by the end user. Additionally, the data between the institutional networks and students must be easy transmission and reproduction.

It is also possible to create notes in the languages of users and in its alphabet. The GIDEON has 4 modules: diagnosis, epidemiological, therapy and microbiology. The module tool is responsible for generating a ranking of different diagnoses based on signs, symptoms, clinical analyzes, incubation period and country of origin of the disease. This module is a list of differential diagnosis that enables the reader of this report, which can be printed or sent to an email, access a table comparing the clinical characteristics of the diseases listed and issues related to the omission or classification of specific diseases. The module of epidemiology has data of many infectious diseases generic but also specific to

certain countries. All data in the GIDEON come from the Ministry of Health, military agencies, specialized lists that are on the Internet, scientific articles and data presented in major conferences.

The EPIPOI is a tool user-friendly interface that allows the exploration and extraction of parameters describing trends, seasonality and defects that characterize the epidemiological diseases [4]. It is also possible to visualize and explore data for time series that are fundamental in epidemiological analysis, not forgetting the inspection of data by geographical regions. In terms of data visualization, one of the characteristics of EPIPOI is the ease of draw the parameters of the time series extracted via scatterplot. With multiple time series, compare these parameters can be insightful, especially for data with geographical references, with unique data to latitude and longitude. With geographical information, the data can be visualized through maps, where medium-sized, amplitudes and times of peak surveys can be plotted to identify geographical trends.

Currently, in Portugal, the only existing system for notification of notifiable diseases is the SINAVE, which will now be improved so that it is possible to analyze and visualize the data of infectious diseases in real time. In Europe there is a system, the Atlas Surveillance of Infectious Diseases [5], which allows you to view and analyze data relating to some communicable diseases, however, only a small number of diseases is represented and the preview does not have data in real time.

## 3. Implementation

To support the eVD Lab was requested access to a replica of the database of SINAVE housed in the SPMS (Shared Services of the Ministry of Health).

### 3.1. The system architecture

The system consists of two main components, the backend and frontend. In backend are obtained the data and are analyzed in SQL and R. In the frontend are generated the views in R with Highcharts library.The system consists of two main components, the backend and frontend. In backend are obtained the data and are analyzed, used SQL and R. In the frontend are generated the views by using the R and The Highcharts library. Since there are no significant changes in real time in the data, the system is prepared to implement caches in both frontend and backend. These caches will store the data analyzed and the views generated after the analysis on the backend and after the creation of views in frontend, respectively. The caches will perform and save data for a period of 10 minutes (Figure 1).

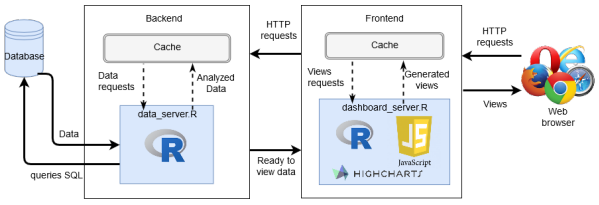


Figure 1: The system architecture.

### 3.2. Data

In this database it was obtained information from the clinical laboratory as name, date of birth, gender and place of residence of the patient and even the disease analyzed, date of collection, date of completion and results of analysis. For this it was used SQL queries.

Clinical laboratories can notify through two ways interoperability and webservice. On the basis of data these data are stored in their tables so it was necessary to deal with them. This process has been programmed in R, joining the data on a single date frame. One of the requirements that the system should fulfill was the allocation of date of notification, in which the final date assigned by the system should be the least of three cases: harvest date, date of completion and date of entry of the notification in SINAVE. Once all these data are obtained directly from the database, this analysis is incorporated in the query, taking advantage of the capabilities of SQL in calculating the lowest of these three fields, where they exist.

Another of the requirements that the system should fulfill was the identification of duplicate cases, just considering them only once. There are diseases that are not curable today, such as the hepatitis C and HIV, so that in the event of a notification for a particular person with this disease at SINAVE, the system should not consider all analysis that that same person come to realize and accuse positive, as well, will make its entry in the system cases that are no more than a repetition of a case already reported. However, it is important to stress that laboratories should report all cases to the SINAVE to the DGS has all the information that exists only in the new system developed for the purpose of analysis, are not considered to be duplicated, because it is a piece of information that will be publicly visible and if they were considered all cases would be wrong under the epidemiological point of view.

o ensure that the cases were duplicates detected by the system, it was necessary to obtain the name of the persons listed in the notification to detect which in fact are duplicated, and are subsequently deleted the name of the person for privacy issues. How can occur if two or more people share the

name, it was a combination of the field the name of the person with the date of birth, since the probability of two or more people share the name and date of birth is more limited. After the junction of the name with the date of birth, only if you considered the notification date of notification, thus ensuring a greater accuracy on the first date that the person accused in a positive laboratory test. However, this process can only detect and filter the cases of diseases that have no cure, i.e., should only be considered a first date that the result of the analysis is positive since it is not curable. All laboratory tests subsequent to that date will have a positive result, but there are diseases that have healing only during a certain period of time, if the person perform new clinical analyzes, and that is a good thing, should not be considered as new cases are still considering concerning the previous case to be temporally close. This period in which it is to be regarded as concerns the same case previously reported, considering it a duplicate, varies depending on the disease, being that there are diseases whose period is one year and diseases in which the period is five years. Initially was saved with a list of the diseases that should be considered new cases only after one year of first date that occurs a laboratory tests which is positive, another list with diseases that should only consider new cases after 5 years since the first day of a positive analysis, another list with diseases that should only be considered the first notification and a new list with the other diseases that should be regarded as new cases every new notification in SINAVE. In this role, it is checked that list is the disease of each notification and are returned the notifications that meet the temporal interval. Depending on the list you are in, you attribute an interval in days, whereas a year corresponds to 365 days and 5 years to 1825 days, so you can use the potential of R to calculate the difference in days between two dates. After the award of the gap, are selected all notifications that have as a difference of dates a number greater than the gap, that is, to iterate over each notification, there is a notification of notifications already exist in the system, and if the system already has a notification to that person, and, at the same time, if the difference between this date and the date of the new notification is greater than the gap, meaning that is not duplicated. The date of this new notification shall be deemed to be the most recent date.

It should be noted that in spite of this process of duplicate detection cover many of the cases, there are still considering the cases in which the laboratory tests can be carried out anonymously. In these cases, it is completely impossible to see any kind of duplicates.

For the system to be able to consider the groups pursued, it was necessary to calculate age from date of birth. Like the other dates stored in the database, the date of birth, it is with the format date, but as a text string. And there was a need to convert to date. However, it was found another problem. The date of birth is stored in the database has only two characters for the year, being stored in the format DD-MM-YY, and DD, days from 1 to 31, MM, month 1 to 12 and YY the last two digits of the year. This causes there is uncertainty in determining the year of birth, for example, someone born to "15-06-17", it is not known whether concerns a person who was born in the year of 1917 or in 2017. It was considered that all dates are later than the current date, the characters of the year relate to years of 1900's and not 2000's. This makes you raise another issue. The system will not exist people with more than 99 years, in cases in which a person has more than this age, the system will consider less 100 years to the age that has, for example, a person with 104 years, will be considered as having 4. However, it was the best solution, because, compared to the number of centenarians in Portugal, would have little impact on the analysis. From the date of birth was calculated the age being later categorized in the following groups 0-1, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74 and 75 years.

In order to fulfill the requirement of the system show information about the location of the occurrence of reported cases, it was necessary to obtain the place of residence of the person concerned in the notification. This information is obtained, in SQL query , through the field of the parish of the wearer of the tables or data derived directly from the portal of SINAVE whether data derived by interoperability. This field contains a code. This code is on the consolidation of locations designed by INE (National Institute of Statistics), which is a unique code and through which you can get to town, county, district, NUTS I, NUTS II and NUTS III once the code is composed of nine digits in the format, "ABCxxyyzz", where "A" concerns the NUTS I, "AB" to the NUTS II, "ABC" the NUTS III, "xx" to the district, "xxyy" to the county and "xxyyzz" to the parish.

In relation to the requirement to show information by gender, the data are not stored in the database in the same way, i.e., the data from SINAVE are not represented in the same way that the coming of interoperability. The genus in the data of the portal of SINAVE are represented by "F", "M" and "N", corresponding to foodstuffs female, male and unknown, respectively, while the data interoperability are represented by the same order, by "18", "17" and "N" . In order to gather the data and be consistent, that all data represented by "F" and "18' to designate by "Female", "M" and

"17", with "Male" and "N' and "0", which correspond to cases in which there is no information in the genre, referred to as the "unknown'.

### 3.3. Visualization

In the frontend are generated the views. It was used a dashboard with a side menu and an information panel. With the dashboard information is centralized and structured and easier for the user to search what you want (Figure 2). To represent the data for the age group were used bar charts using the ideology of multi-panel [6] where can aggregate more than one information in an only view, in this case age and gender (Figure 3). One of the


Figure 2: eVD Lab dashboard.

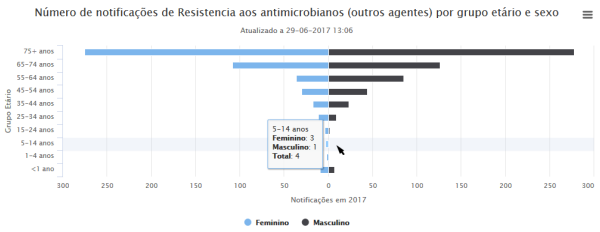


Figure 3: Chart to show age group and sex.

menus it is possible to see the notifications per day of each disease, with a panel which shows the total number of cases on the day, week, month and year (Figure 4). The notifications per day are available through a line graph (Figure 5). To see the intensity with which each disease is notified by day was created a heatmap where darker colors represent the largest number of notifications.To see the intensity with which each disease is notified by day was created a heatmap where darker colors represent the largest number of notifications (Figure 6). As the axis are represented every day the zoom feature is available. In order to show the location of the incidence of laboratory notifications of notifiable diseases in Portugal, were used choropleth maps. This maps you can see the incidence by NUTS III, ARS, district and county per month, however are not broken down by disease, is only represented the total

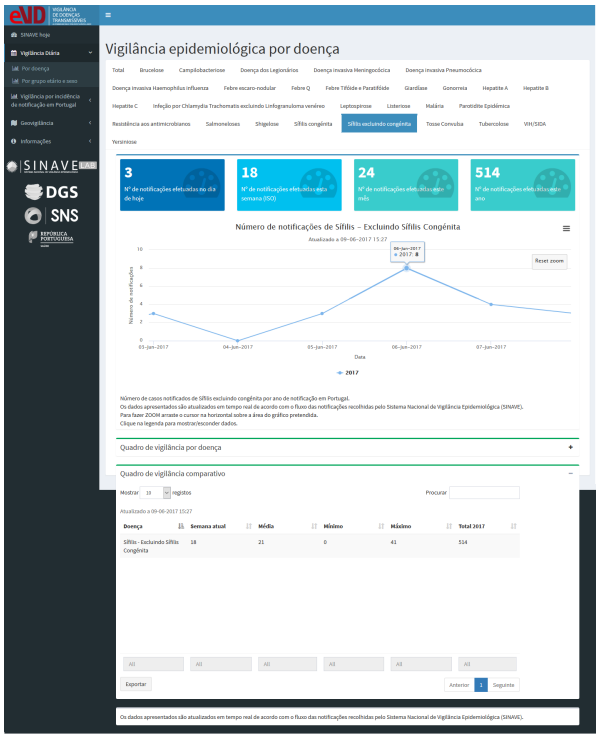


Figure 4: Disease information.

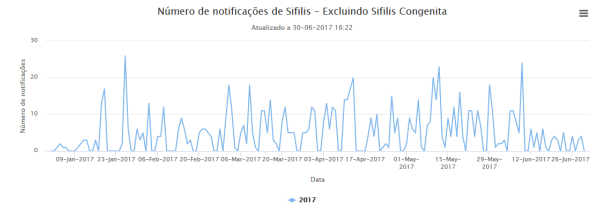


Figure 5: Line chart to show notifications per day of selected disease.

number of notifications. In map is still possible to choose the month view through a timeline, dragging the cursor over the same. If you want to view in an automatic way the passage of months, you can also do this by simply that actuate the play button. The map shows the number of notifications of notifiable diseases in Portugal, by 100,000 inhabitants. That is why maps are accompanied by tables with information notifications of each disease being organized by day, month and year. It is thus possible to analyze disease by disease and it is possible to export the data for further analysis (Figure 7). The tables have filters on all columns so that it is easier to look for the disease, day/month/year and/or region.

## 4. Results

The assessment consisted of two components: user tests and case studies. The tests with users had as objective to evaluate the system the level of in-

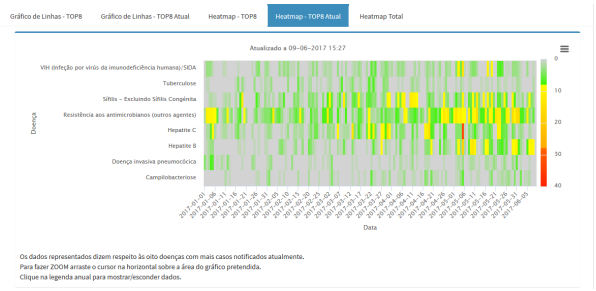


Figure 6: Heatmap to show intensity of notification of diseases per day.

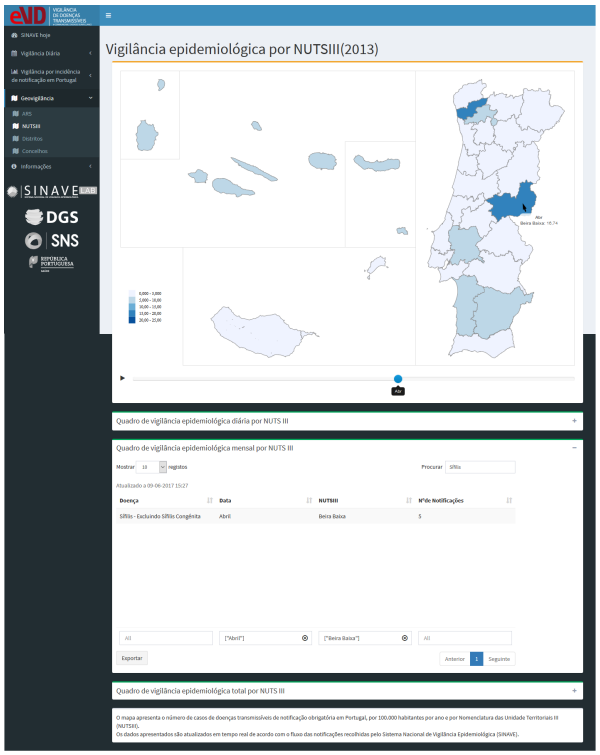


Figure 7: Choropleth map and monthly table.

teractivity and usability, being the users volunteers without knowledge of the field. Test scenarios the objective was the assessment by potential users of the DGS, simulating tasks that they do today without the system, realizing it was functional, intuitive and with good usability.

### 4.1. User tests

In order to see if the system allows you to collect accurate information about the current state of epidemiological surveillance in Portugal, such as the number of cases of each compulsorily notifiable disease, age group, gender and region of patients and whether it is easy to use, perceptible, attractive and pleasing to the user, tests were carried out with users. Being the target audience of the system, the general public, the system was tested by volunteers.

The test with users were conducted with 20 volunteers between 12 and 26 June, taking place in the technical room of the Department of Computer Engineering 0.07, the flag of Informatics III, on the campus of the Alameda do Instituto Superior Tcnico and focused primarily on five tasks. The tasks exercised the various parts of the system, covering all the features required when defining requirements and was not necessary for the user to carry any material for a test session.

The five tasks, which were made by users in a random order, are:

1. Know the number of notifications of SINAVElab today.
2. Know the age group with a higher incidence of Hepatitis B.
3. know what the disease with more notifications in the SINAVElab on May 5, 2017.
4. Find out which is the ARS with the highest incidence of notifiable diseases in May and which disease with the most notifications for that month for this ARS.
5. Know if this week, the disease with the most reports of NUTSIII with the highest incidence of notifiable diseases in February, is above or below the average of the weeks of the year.

The test session was divided into four phases:

1. The first one, lasting 5 minutes, consisted of a brief explanation of the test system and scope of the test. It was also reminded to the user that he would test the system, that it was not the user that was going to be tested, but the system, not having to be embarrassed or afraid of making mistakes.
2. The user was then given full freedom to freely explore and use the system for 5 minutes.
3. After a first contact with the system, the user performed the set of 5 tasks in random order, thus ensuring intra-task independence. This phase had the time necessary for the user to complete the 5 tasks, and it was initially expected that this phase had a maximum duration between 10 and 15 minutes, which has been proven.
4. After completing the tasks, the user was asked to complete a satisfaction questionnaire to assess some aspects of the evaluated system. This phase lasted for 5 minutes.

The test phase consisted in random assignment of the tasks to be performed, asking the user what

order they wanted to perform the tasks. Before starting any of the tasks, the starting point was the main page of the system, menu "SINAVE today". Each task was considered to have a correct end when the user rightly responded to the question implicit in the task, not having any help in its accomplishment, except if the user "block", remaining a long time stopped.

#### 4.1.1 The users

The 20 users who tested the system were volunteers, divided into two age groups: 90% of users between 18 and 30 and 10% of users between the ages of 41 and 60, mostly Male, 70 % of users. In terms of the degree of complete instruction, users are mainly distributed through High school, 50 % of users and Bachelor's degree, 45 % of users.

#### 4.1.2 Data collected

During the tests performed by the user, data were collected such as the time for each task as well as the number of errors, and it was considered an error when the user walked through an unnecessary table, graph or menu to perform the task. The collected data represented in the table 1.

| | Tasks | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| User | 1 | | 2 | | 3 | | 4 | | 5 | |
| 1 | 11 | 0 | 45 | 0 | 45 | 2 | 53 | 0 | 300 | 4 |
| 2 | 24 | 1 | 103 | 1 | 55 | 3 | 92 | 1 | 355 | 4 |
| 3 | 19 | 0 | 195 | 2 | 31 | 0 | 74 | 1 | 621 | 7 |
| 4 | 1 | 0 | 163 | 0 | 35 | 0 | 44 | 0 | 344 | 2 |
| 5 | 1 | 0 | 26 | 0 | 112 | 3 | 93 | 0 | 363 | 4 |
| 6 | 2 | 0 | 30 | 0 | 64 | 1 | 51 | 0 | 222 | 2 |
| 7 | 62 | 1 | 35 | 1 | 18 | 0 | 78 | 0 | 77 | 1 |
| 8 | 25 | 0 | 60 | 0 | 25 | 0 | 51 | 0 | 100 | 0 |
| 9 | 23 | 0 | 56 | 0 | 53 | 0 | 55 | 0 | 104 | 0 |
| 10 | 1 | 0 | 23 | 0 | 90 | 1 | 60 | 0 | 134 | 0 |
| 11 | 1 | 0 | 31 | 0 | 39 | 0 | 64 | 0 | 184 | 0 |
| 12 | 1 | 0 | 53 | 0 | 26 | 0 | 88 | 0 | 245 | 0 |
| 13 | 1 | 0 | 25 | 0 | 19 | 0 | 53 | 0 | 64 | 0 |
| 14 | 2 | 0 | 41 | 0 | 36 | 0 | 58 | 0 | 97 | 0 |
| 15 | 7 | 0 | 31 | 0 | 62 | 0 | 64 | 0 | 71 | 0 |
| 16 | 7 | 0 | 48 | 0 | 71 | 0 | 69 | 0 | 74 | 0 |
| 17 | 22 | 1 | 73 | 0 | 47 | 0 | 102 | 0 | 147 | 0 |
| 18 | 1 | 0 | 22 | 0 | 54 | 0 | 55 | 0 | 132 | 0 |
| 19 | 7 | 0 | 50 | 1 | 98 | 0 | 112 | 0 | 156 | 0 |
| 20 | 1 | 0 | 19 | 0 | 36 | 0 | 62 | 0 | 82 | 0 |
| **Mean** | 10,95 | 0,15 | 56,45 | 0,25 | 50,8 | 0,50 | 68,9 | 0,10 | 193,6 | 1,20 |

Table 1: Times, in seconds, and number of errors committed per user in each of the tasks performed in the test session.

Looking at the table 1 notes that the average time to carry out the tasks increases from first to last, that expected, since the complexity of the same is also growing, being the fifth task more complex. In the first task, which had a lower degree of complexity had as average 10.95 seconds. To carry out this task was only necessary to observe the information panel of the home page, and it is the answer without any interaction being visible in a panel with the value of the response. In this sense, the average time is within the expected range, because users only had to look for the location of information. Considering the number of errors committed, an average of 0.15, is also within the expected given the simplicity of the task. The mistakes made consisted in finding exactly the same answer but another less menu immediately. The Tasks 2, 3 and 4, have similar complexity, requiring the user to navigate through the menus in the system. The average times are similar, 56.45; 50.8 and 68.9 seconds for the tasks 2.3 and 4 respectively. These times can be considered as good as required navigation between menus, and interpretation of the bar graph with need of interaction to obtain in tooltip the answer in Task 2, as well as interpretation of a map and consequent reading table filtering and comparing the values it presented in tasks 3 and 4. The number of errors is also positive because in none of the tasks the average number is greater than 0.50, and consisted of the erroneous reading of the age at tooltip of bar graph in task 2 and in the analysis of daily table instead of monthly tables in tasks 3 and 4. The last task required the navigation at a first menu with interpretation of a map and subsequent reading, filtering and comparison of values in the table monthly to get the first part of the answer. Getting this response, the disease with higher incidence in the month of February in NUTSIII with largest number of notifications, required even the navigation in another menu and attentive analysis of a table with comparative values. This task took, on average, 193.6 seconds, which equates to 3 minutes and 14 seconds. This value is positive because the task requires interaction with the system and the response is divided into two phases, recalling the little time of prior use of the system by users who have tested and is expected an improvement to the extent that the use is more common. The average number of errors was 1.20, which is a good thing because it involved the navigation between various menus and consultation of a map and tables, consisting mostly of reading of daily tables instead of monthly. Considering the global values, it can be assumed that the system had a good result because the number of errors is low, indicating that the system is intuitive. And yet, in less than five minutes it is possible to obtain information, without the need to obtain any authorisation by the DGS, which currently can take days.

### 4.2. Satisfaction questionnaire

The satisfaction questionnaire consisted of a questionnaire of 15 multiple-choice questions, being the possible answers given on a scale of 1 to 5, where 1 corresponds to "Disagree" and 5 corresponds to "strongly agree", and yet 1 question to answer open. Of the 15 multiple-choice questions, 10 relate to the assessment by the user of the statements of the SUS

(Scale of Usability of Systems) and which are:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The answers are then converted to a score of SUS

Through the average score SUS you can compare with the score which is considered standard for a good system in terms of usability, and this score of 68 points. Systems that have score from SUS average below 68 points shall be considered systems that have to be strongly improved the level of usability, while above 68 points shall be considered systems that have an usability is above average. The score reference pattern can be more accurate, considering that systems with the average scores above 74 have usability very good and are pleasing to the user. The average scores above 80.3 are considered of great usability, being fully pleasing to the user, recommending the system. The evaluation on the part of users already converted into score SUS, is shown in Table 2. The mean score SUS is 77,375

| | Questions | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| User | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total | Total x 2.5 |
| 1 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 2 | 36 | 90 |
| 2 | 1 | 3 | 3 | 4 | 3 | 3 | 3 | 2 | 2 | 0 | 24 | 60 |
| 3 | 3 | 2 | 2 | 1 | 3 | 3 | 3 | 3 | 2 | 0 | 22 | 55 |
| 4 | 1 | 3 | 3 | 4 | 3 | 4 | 3 | 4 | 3 | 4 | 32 | 80 |
| 5 | 2 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 4 | 3 | 35 | 87,5 |
| 6 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 3 | 2 | 1 | 20 | 50 |
| 7 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 38 | 95 |
| 8 | 4 | 4 | 3 | 4 | 4 | 4 | 1 | 3 | 2 | 3 | 32 | 80 |
| 9 | 3 | 4 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 32 | 80 |
| 10 | 3 | 3 | 3 | 2 | 3 | 4 | 3 | 4 | 3 | 1 | 29 | 72,5 |
| 11 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 3 | 4 | 4 | 38 | 95 |
| 12 | 3 | 3 | 3 | 2 | 4 | 3 | 2 | 2 | 2 | 2 | 26 | 65 |
| 13 | 3 | 4 | 4 | 4 | 4 | 3 | 3 | 4 | 3 | 1 | 33 | 82,5 |
| 14 | 3 | 3 | 3 | 4 | 4 | 4 | 2 | 3 | 3 | 4 | 33 | 82,5 |
| 15 | 4 | 1 | 3 | 1 | 3 | 3 | 3 | 3 | 0 | 2 | 23 | 57,5 |
| 16 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 40 | 100 |
| 17 | 3 | 3 | 3 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 34 | 85 |
| 18 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 4 | 32 | 80 |
| 19 | 2 | 4 | 3 | 1 | 3 | 2 | 3 | 4 | 3 | 1 | 26 | 65 |
| 20 | 3 | 3 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 2 | 34 | 85 |
| **Mean** | 2,8 | 3,3 | 3,15 | 3,05 | 3,5 | 3,45 | 3,1 | 3,4 | 2,85 | 2,35 | 30,95 | 77,375 |

Table 2: SUS score of 10 questions.

points, which means that it is above the average of 68 and therefore cannot be considered that the system has a good usability. When compared with the standard score of more accurate reference, it falls short of the 80.3, value from which you consider a excellent usability, however is a high value, allowing the measurement that are needed few changes to make the usability of the system excellent. In questions 1, 9 and 10, the average score is less than 3. If we consider the content of these issues, we note that can be interconnected as they approach the trust (question 9), the necessity of learning (issue 10) and the frequency of the use of the system (question 1). A possible conclusion that can draw is that the fact that the users feel that they do not have sufficient knowledge of the domain of the system makes them feel less confident in their use and feel the need to learn before you use it. To improve these aspects should make the system more intuitive and appealing during the description of certain technical terms for which the user will feel more confident and don't feel the need to learn how to use it, thus providing a greater frequency of use.

### 4.3. Case studies

In addition to the tests with volunteer users, the system was also subject to an evaluation by two technical colleges of DGS who will use the system. This test consisted in simulating navigation tasks that will make the use of the system. The process of development of the task consisted in carrying out the task by user, following the approach of saying out loud what you are thinking and what you want to do, commenting on the interactions that went with the system.

One of the tasks was the identification of the number of measles cases occurred in Portugal in 2017. The main menu, "SINAVE today", both users found unintuitive the division of diseases, that is, the diseases with low incidence diseases with less than 30 cases over the last 4 years, are separated from the rest of the notifiable diseases, but a user with little knowledge of the domain does not know the distinction of which diseases, browsing through it to the menu of diseases with more notifications that is in the menu "Daily Surveillance" . To make this clearer division, was suggested by users who add a caption to explain the division of diseases in the menu of diseases with more notifications. This division exists because of the difference in representation of the data, because the diseases with low incidence have few cases, it would not be appropriate to use the same approach of other diseases, which priori, have many more cases. Also commented that, although the point mentioned above, the system was well done.

The other task that users have chosen to per-

form was the identification of age groups and sex of various diseases. In this task the reviews were positive, being that the navigation was intuitive and the graphic was helpful and very noticeable.

In the remaining navigation is to emphasize that it was referred to the fact that the maps add little information, since it was more useful to them a map where you could see the geographical distribution of the disease instead of seeing the incidence without discrimination of diseases. However, this case does not appear in the requirements initially defined, so that was not the target of development, combining the fact that once more, not wanting to show the general public all details possible not to alarm the population.

### 4.4. Discussion

Analyzing the results of the tests and the list of requirements, it can be assumed that the system was well done and that objectives were achieved. The session of tests with users shows that the system allows quick perception of the data on the epidemiological surveillance in Portugal and that the interaction is intuitive, since the tests performed the number of errors is low, not forgetting the fact that the user only interacted with the system in the session itself, what perspective that when returning to use the system will have better performance, and even the average time of performance shall be appropriate to the complexity of the tasks required. It is also possible to conclude that the system has a good usability, reaching to average 77,375 points in the SUS, getting above the average of 68 points of reference and getting slightly below the average of 80.3 points from which systems are considered as having excellent usability.

Through the tests with users of DGS Notes that the analysis performed automatically and present in the system, is functional and covers all points requested, by complying with the requirements originally proposed, showing that they were still pleased with the developed system.

## 5. Conclusions

The system developed, the eVD Lab, had as its main objective to provide a means of surveillance of notifiable diseases in Portugal, in real time. This goal was achieved, and it is now possible to use the eVD Lab to obtain real-time information about the current state of laboratory notifications of notifiable diseases in the United States.

In terms of the requirements, were all achieved, since it is possible to view in real-time information about the incidence of notifiable diseases by several factors such as geography, age groups and gender. The system also allows you to see trends and developments over time. It is also possible to export the data for analysis in the DGS and or by external entities. The layout of menus and the visual elements such as maps, heatmaps, line graph and tables available in eVD Lab were decided in conjunction with the staff of the division of epidemiology and surveillance of DGS the responsibility of Dr. Ctia Sousa Pinto, based on his experience of what is most relevant and necessary to the general public and health professionals who potentially will use the system. In addition to the requirements and objectives proposed, were also added the following features:

- Know the 8 mandatory reporting diseases with the most notifications since the beginning of the year in real time.
- Know the number of laboratory reports of each disease, except those of low incidence, broken down by day, week, month and current year.
- Compare the number of laboratory reports, broken down by disease, from the current week to the weekly average.

Besides the requirements have been fulfilled, through the evaluation with users it is verified that the system has a good usability, ensuring a good interaction of users with it.

### 5.1. Contribution

With eVD Lab it is now possible to view a variety of real-time information on notification of notifiable diseases in Portugal, a task that until the existence of this system was not possible. The development of this innovative system contributes to the modernization of the means of control and prevention of epidemiological diseases, short-term objectives of the CDC and the ECDC, placing Portugal as one of the pioneers in the public provision of a data visualization system on mandatory reporting diseases.

### 5.2. Future work

The developed system, the eVD Lab, is only the first version, so it can be improved in several aspects in the future. By acquiring new knowledge about disease prevention, eVD Lab can and should accompany this evolution by adding new analyzes to the data, ensuring that the system remains up-to-date, accurate and objective in the analysis.

In new versions of eVD Lab, new visualizations can also be added to ensure that the system can collect new information that is more detailed and detailed than the current ones.

The current version was developed with the general public in mind, and it is not necessary to have a great knowledge of the domain in order to understand the data presented. However, in my opinion, it would be very useful for DGS to have a private version of eDD Lab. Which could be considered the possibility of having an authentication system that

allows a presentation of the data private so that the DGS can obtain more in-depth analyzes on the SINAVE Lab data, analyzing them in more detail and paying attention to the attributes of the notifications made Laboratory clinics. In this private version, which would be accessed through authentication, could also have more accurate and detailed views that can not be publicly disclosed, such as the geographical distribution at parish level, and also the same geographic distribution but discriminated by disease.

Another point that can be improved in future versions of eVD Lab is the technology used, especially in the visualization layer. This layer, when developed in R, although corresponding to the intended one, limited the potential of visualizations used. It could be considered to improve this layer using JavaScript because it allows greater freedom in creating visualizations, allowing for greater interactivity in the visualization itself and between visualizations, something that with the R and the Highcharts library did not exist.

**Acknowledgements**

First of all I would like to thank my family members for all the support they have given me throughout my academic career and without which this project would not be possible.

I would also like to thank my supervisor Dr Ctia Sousa Pinto for giving me the opportunity to enter an internship within the scope of the DGS epidemiology division and for all guidance during the course. Not forgetting also the whole team of the division of epidemiology with whom I shared many moments in the DGS, namely Jos Loff, Paula Vicncio, Clia Gaspar, Maria Joo and Lurdes Morgado, and two other trainees with whom I had many experiences, Daniela Pimentel and Francisco Duarte.

I would also like to thank the supervisors Professor Daniel Gonalves and Professor Mrio Gaspar for all the guidance and support provided, advice, availability and sharing of knowledge that without which this thesis would not be possible.

Last but not least, I would like to thank all my friends and colleagues for all the times shared, the good ones and the less good ones, and without them it would not be the same thing. I would like to add that all the moments of pressure that we have passed along the course are worthwhile, because with them we have evolved and acquired new capacities that will surely add value in the near future.

To each and every one of you, thank you very much.

**References**

[1] Cody Dunne, Michael Muller, Nicola Perra, and Mauro Martino. VoroGraph: Visualization Tools for Epidemic Analysis. *Extended Abstracts of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 2:255–258, 2015.

[2] Yarden Livnat, Theresa Marie Rhyne, and Matthew Samore. Epinome: A visual-analytics workbench for epidemiology data. *IEEE Computer Graphics and Applications*, 32(2):89–95, 2012.

[3] Stephen C. Edberg. Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 40(1):123–6, jan 2005.

[4] Wladimir J Alonso and Benjamin J J McCormick. EPIPOI: a user-friendly analytical tool for the extraction and visualization of temporal parameters from epidemiological time series. *BMC public health*, 12(1):982, 2012.

[5] Surveillance atlas of infectious diseases. `http://ecdc.europa.eu/en/surveillance-atlas-infectious-diseases`. Accessed: 2017/07/14.

[6] Kenneth K H Chui, Julia B. Wenger, Steven A. Cohen, and Elena N. Naumova. Visual analytics for epidemiologists: Understanding the interactions between age, time, and disease with multi-panel graphs. *PLoS ONE*, 6(2), 2011.