J Med Syst (2017) 41: 51
DOI 10.1007/s10916-017-0697-4

Data Mining in HIV-AIDS Surveillance System

Application to Portuguese Data

Alexandra Oliveira[1,2,3] · **Brígida Mónica Faria**[2,3] · **A. Rita Gaio**[1,4] · **Luís Paulo Reis**[2,5]

Received: 21 June 2016 / Accepted: 5 February 2017 / Published online: 18 February 2017

Abstract The Human Immunodeficiency Virus (HIV) is an infectious agent that attacks the immune system cells. Without a strong immune system, the body becomes very susceptible to serious life threatening opportunistic diseases. In spite of the great progresses on medication and prevention over the last years, HIV infection continues to be a major global public health issue, having claimed more than 36 million lives over the last 35 years since the recognition of the disease. Monitoring, through registries, of HIV-AIDS cases is vital to assess general health care needs and to support long-term health-policy control planning. Surveillance systems are therefore established in almost all developed countries. Typically, this is a complex system depending on several stakeholders, such as health care providers, the general population and laboratories, which challenges an efficient and effective reporting of diagnosed cases. One issue that often arises is the administrative delay in reports of diagnosed cases. This paper aims to identify the main factors influencing reporting delays of HIV-AIDS cases within the portuguese surveillance system. The used methodologies included multilayer artificial neural networks (MLP), naive bayesian classifiers (NB), support vector machines (SVM) and the k-nearest neighbor algorithm (KNN). The highest classification accuracy, precision and recall were obtained for MLP and the results suggested homogeneous administrative and clinical practices within the reporting process. Guidelines for reductions of the delays should therefore be developed nationwise and transversally to all stakeholders.

Keywords Data mining · Surveillance system · Surveillance data · HIV-AIDS · Reporting delay

This article is part of the Topical Collection on *Systems-Level Quality Improvement.*

✉ Alexandra Oliveira
aao@ess.ipp.pt

Brígida Mónica Faria
btf@ess.ipp.pt

A. Rita Gaio
argaio@fc.up.pt

Luís Paulo Reis
lpreis@dsi.uminho.pt

[1] Center of Mathematics, University of Porto, Porto, Portugal

[2] Artificial Intelligence and Computer Science Laboratory, LIACC, Porto, Portugal

[3] ESS-IPP - Higher School of Health, Polytechnic of Porto, Porto, Portugal

[4] Department of Mathematics, Faculty of Sciences, University of Porto, Porto, Portugal

[5] DSI-EEUM - Information Systems Department, School of Engineering, University of Minho, Braga, Portugal

Introduction

The Human Immunodeficiency Virus (HIV) is an infectious agent that attacks the immune system cells. Without a strong immune system, the body becomes very susceptible to serious life threatening opportunistic diseases. Generally, soon after the initial HIV infection there is an accute illness, with very few specific symptoms, and then the infected individual becomes asymptomatic (A) usually for several years; when diseases start to appear it is said that the individual is in a Symptomatic Condition (SC) and when severe symptoms emerge, it is said that the individual has Acquired Immunodeficiency Syndrome(AIDS).

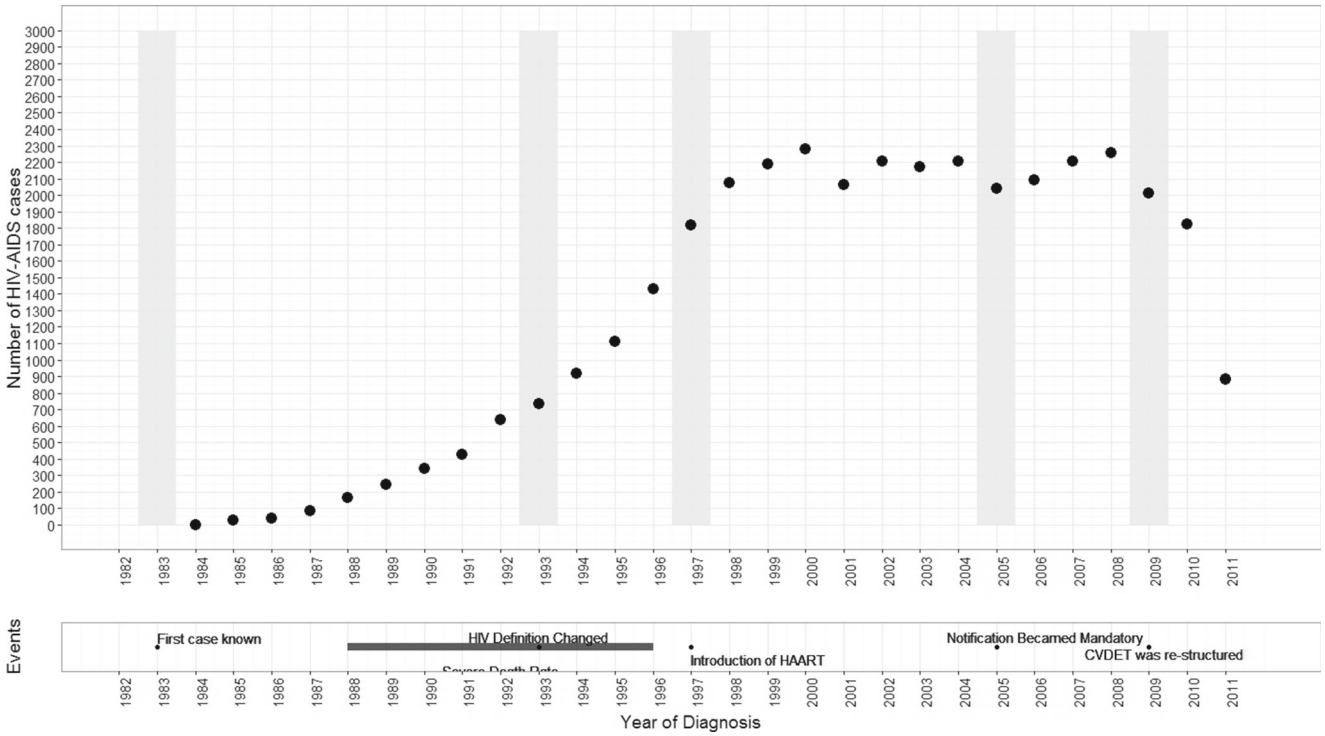

Fig. 1 Number of HIV-AIDS cases diagnosed in Portugal per year of diagnosis. Historical events within the surveillance system are highlighted

There are key populations at higher risk of HIV exposure (injection-drug users (IDU), men who have sex with men (MSM), female sex workers, clients of sex workers ....) and that has to do with the disease transmission mode [1, 2]. Exchange of certain body fluids from infected individuals is necessary but not enough. These fluids must be in contact with a mucous membrane or damaged tissue or be directly injected into the bloodstream [3, 4].

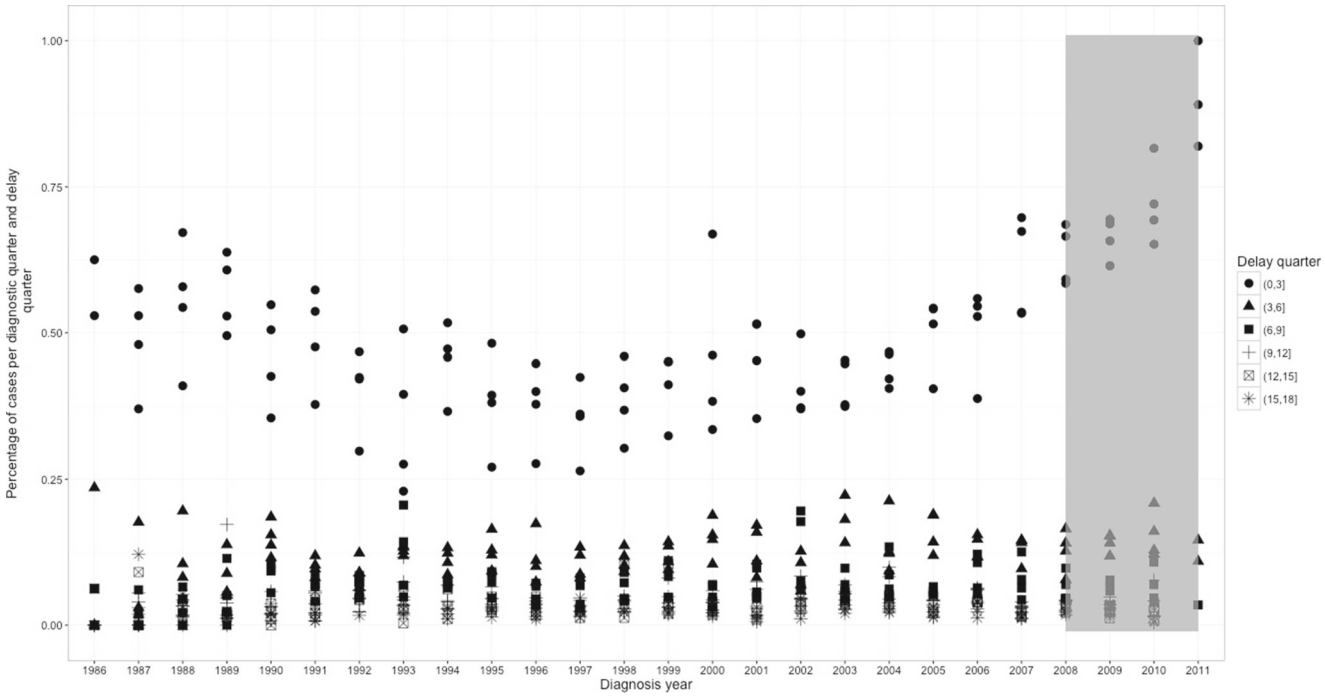

Fig. 2 Percentage of HIV-AIDS cases per diagnosis and delay quarters. The *lighted region* identifies those (recent) years that have to be corrected

In spite of the great progresses on medication and prevention over the last years, HIV infection continues to be a major global public health issue, having claimed more than 36 million lives over the last 35 years since the recognition of the disease [5, 6]. Geographically, the burden of the disease is not equally distributed. For instance, 7 out of 10 people worldwide living with HIV are in sub-Saharan Africa, where this infection is a leading cause of death among adults, women on reproductive age and children [6]. In Portugal, the UNAIDS estimated the prevalence (percentage of infected individuals within the population) to be about 0.6 % of the population in 2009. So, HIV-AIDS is an important public health problem.

Monitoring, through registries, of HIV-AIDS cases is vital to assess general health care needs and to support long-term health-policy control planning [7]. Surveillance systems have thus been established to accomplish this critical mission [8].

Typically a Surveillance System depends on several stakeholders, such as health care providers, the general population and laboratories, which challenges an efficient and effective reporting of diagnosed cases. One issue that often arises is the administrative delay in reports of diagnosed cases [9–12]. This phenomenon should always be taken into account in data analysis on numbers from a surveillance system [13, 14].

An administrative reporting delay can be defined as the time mediating from identification of the HIV-AIDS related event and its national reporting [13]. Other studies have mentioned that this problem depends on a number of factors such as geographical region of diagnosis, calendar year, patient age at diagnosis and HIV infection mode [7, 15–19].

The term “data mining” refers to a collection of techniques that provide the necessary actions to retrieve and gather knowledge from an exhaustive assemblage of data and facts [20]. In particular, they can uncover new biomedical and health care knowledge for clinical and administrative decision making as well as generate scientific hypotheses from large experimental data, clinical databases, and / or biomedical literature [21]. Data mining models can be classified into two categories: descriptive (or unsupervised learning) and predictive (or supervised learning) [22]. Descriptive data mining consists of a collection of techniques aiming to discover unknown patterns or relationships in data. This exploratory analysis includes clustering, association, summarization, and sequence discovery [22]. Predictive data mining infers prediction rules from data. It includes tasks such as classification, regression, time series analysis, and prediction [21]. Classification is the most frequently used data mining method with a predominance of the implementation of bayesian classifiers, neural networks, and SVMs (Support Vector Machines) [23].

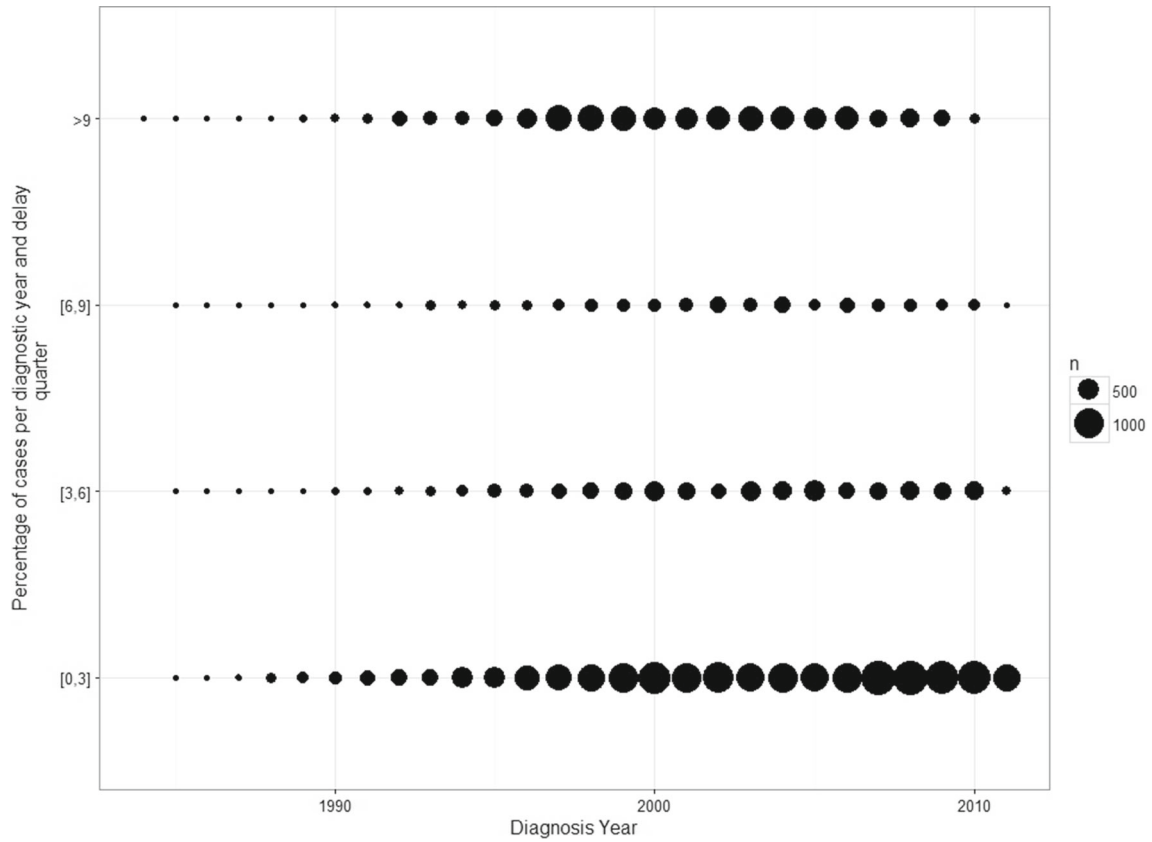

Fig. 3 Number of HIV-AIDS infections by year of diagnosis and reporting delay

This paper aims to identify the main factors influencing reporting delays in the HIV-AIDS cases within the portuguese surveillance system. In order to acomplish this objective, several data mining models were considered, namely multilayer artificial neural networks (MLP), naive bayesian classifiers (NB), support vector machines (SVM) and the k-nearest neighbor and algorithm (KNN).

Related work

Traditionally, the reporting delay distribution has been estimated from the conditional or unconditional log-likelihood of proportional hazards regression models [17–19, 24–30]. Although parametric assumptions allow for the estimation of the distribution of the reporting delays, results are extremely imprecise and depend strongly on the assumptions [8]. Moreover, any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough [31].

As an alternative, nonparametric data mining techniques have been used in cases where similar patient records and related symptoms were used [20].

Portuguese epidemic and data set characteristics

The number of HIV-AIDS cases that has been reported since the 80's places Portugal as one of the most infected countries in Europe [33]. The epidemic is concentrated on high-risk and hard-to-reach sub-populations, thus hindering a timely diagnosis.

Current national directives compel the reporting of all stages of HIV-AIDS along with its progressions and death to the Portuguese Centre for the Transmissible Diseases (CVDET - Centro de Vigilância Epidemiológica das Doenças Transmissíveis), within 48 hours after the event. A proper notification form (in paper format) should be filled in by the patient's medical doctor. However, this procedure has only been mandatory since 2005 [34]. Over the years, the surveillance procedure has suffered some changes that may have altered the quality of the reports. We point out the following:

1. in 1988, the reporting form was altered and more variables were included;
2. in 1993, tuberculosis was included as an AIDS defining disease;

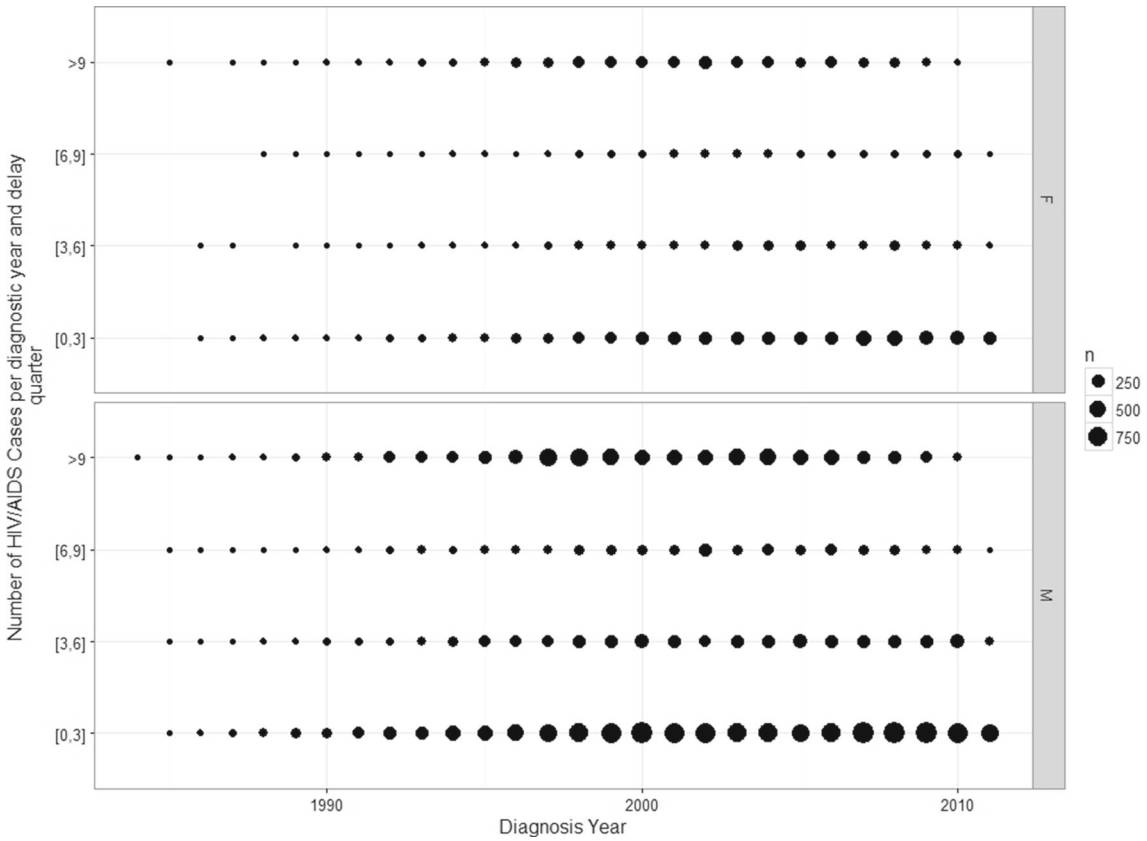

Fig. 4 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and gender (*right vertical boxes*)

3. in 1996, the highly active anti-retroviral therapy (HAART) was introduced;
4. in 2005, the notification forms were re-structured;
5. in 2009, the CVDET was re-structured [14].

The Portuguese HIV-AIDS data analysed within this work consists of nearly 24 attributes from 45000 diagnosed cases, collected between 1983 and 2011. For each patient, it considers the following personal and demographic variables: age at diagnosis, gender, nationality, disease stage, risk group and date of diagnosis. It also contains information on the health providers responsible for the diagnoses and reporting processes.

The analysis considers only HIV 1 virus type. All pediatric cases in children under 6 months of age were excluded. Reporting delays were only considered if lower than 3 years.

Data mining classification in HIV-AIDS Portuguese surveillance data

Data pre-processing

As most HIV-AIDS cases were seen to be reported within the first 3 months after diagnosis, and the reporting behaviour seemed to be homogeneous from that onwards (Fig. 2), reporting delays were discretized into two classes, with a cut-point at 3 months after diagnosis [35, 36]. The authors believe this process also eliminates some of the administrative inaccuracies inherent to the continuous reporting delays. Patients nationalities were classified according to the World Health Organization (WHO) HIV-AIDS Regions.

All residents in Portugal have access to health care services provided by the National Health Service (NHS). It

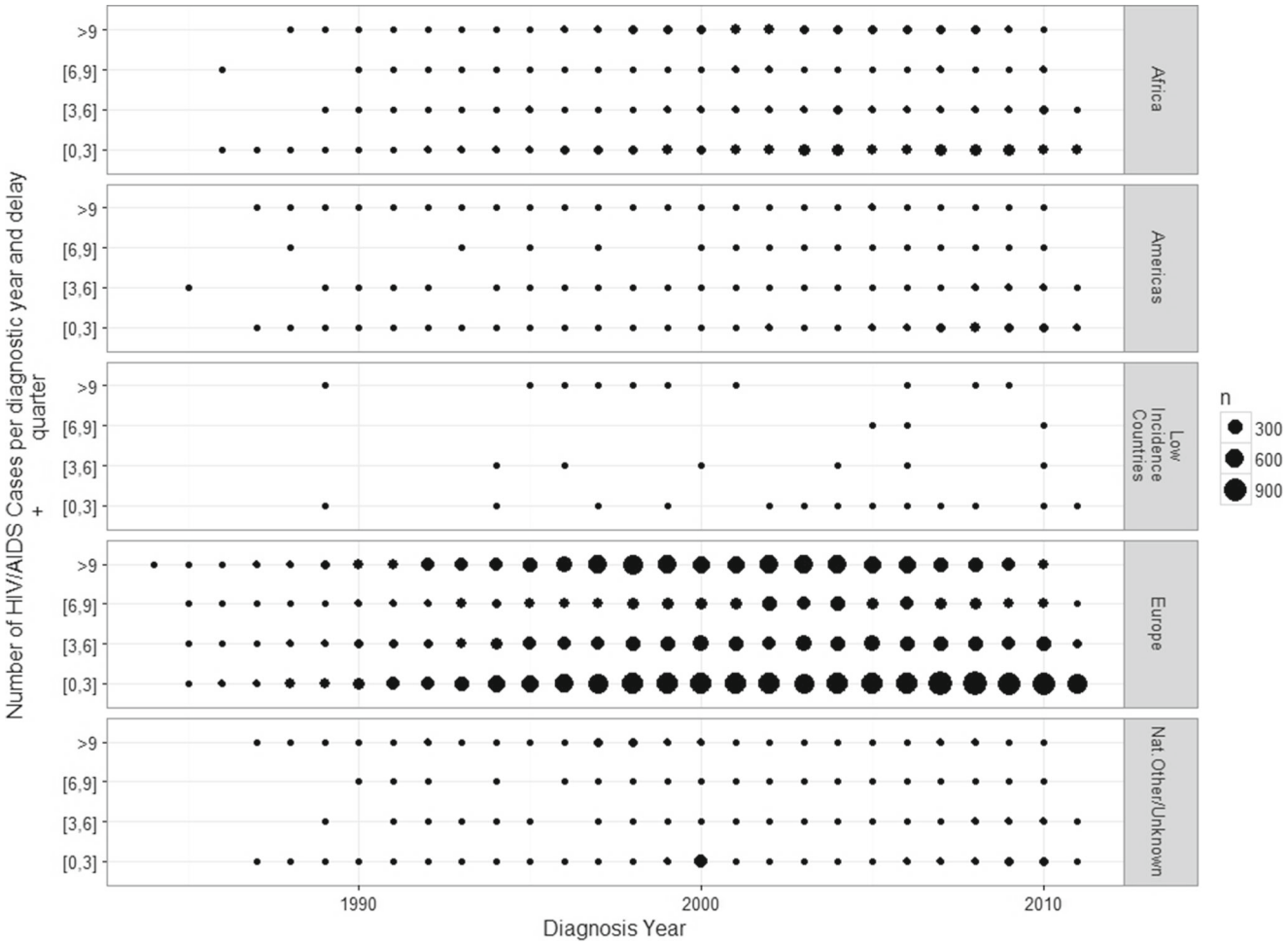

Fig. 5 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and patients nationality (*right vertical boxes*)

is managed at 5 regional levels through Regional Health Administrators (RHA) that are accountable to the Ministry of Health: North, Centre, Lisbon and Vale do Tejo, Alentejo and the Algarve. Each RHA is responsible for the strategic management of its population health, supervision and control of hospitals, management of primary care/NHS primary care centres, centres for treatment of addictive behaviours, and implementation of national health policy objectives [38]. The Ministry of Health cooperates with the Ministry of Justice for providing health care services on prisons and with the Ministry of Defence for providing health care services to the servicemen. Given this structure, the information on the health providers responsible for the diagnoses and reporting processes was cross-classified by type of health care institute and regional administration. Prisons and Military Institutions were considered a single type of health care provider. An additional category ("Admin") was created to accommodate observations without specific health care provider.

Supervised classification was then built upon that 2-class discretization, through feed-forward multilayer perceptron, naive networks, support vector machines and the K-nearest neighbour algorithm with the following input features: age at diagnosis, gender, patient nationality, disease stage, HIV risk group, type of health provider, and administrative and financial responsibility of health care providers.

The predict whether or not a case will ever be reported and to identify which factors influence that characteristic, we used feedfoward multilayer perceptron, naive neural networks, svms and k-nearest neighbour.

Statistical analyses were performed with the R language and software environment for statistical computation, version 2.3.0, and the software plataform RapidMiner 7.1.001.

The models

Feedfoward multilayer perceptron Artificial neural networks are a popular alternative to conventional statistical

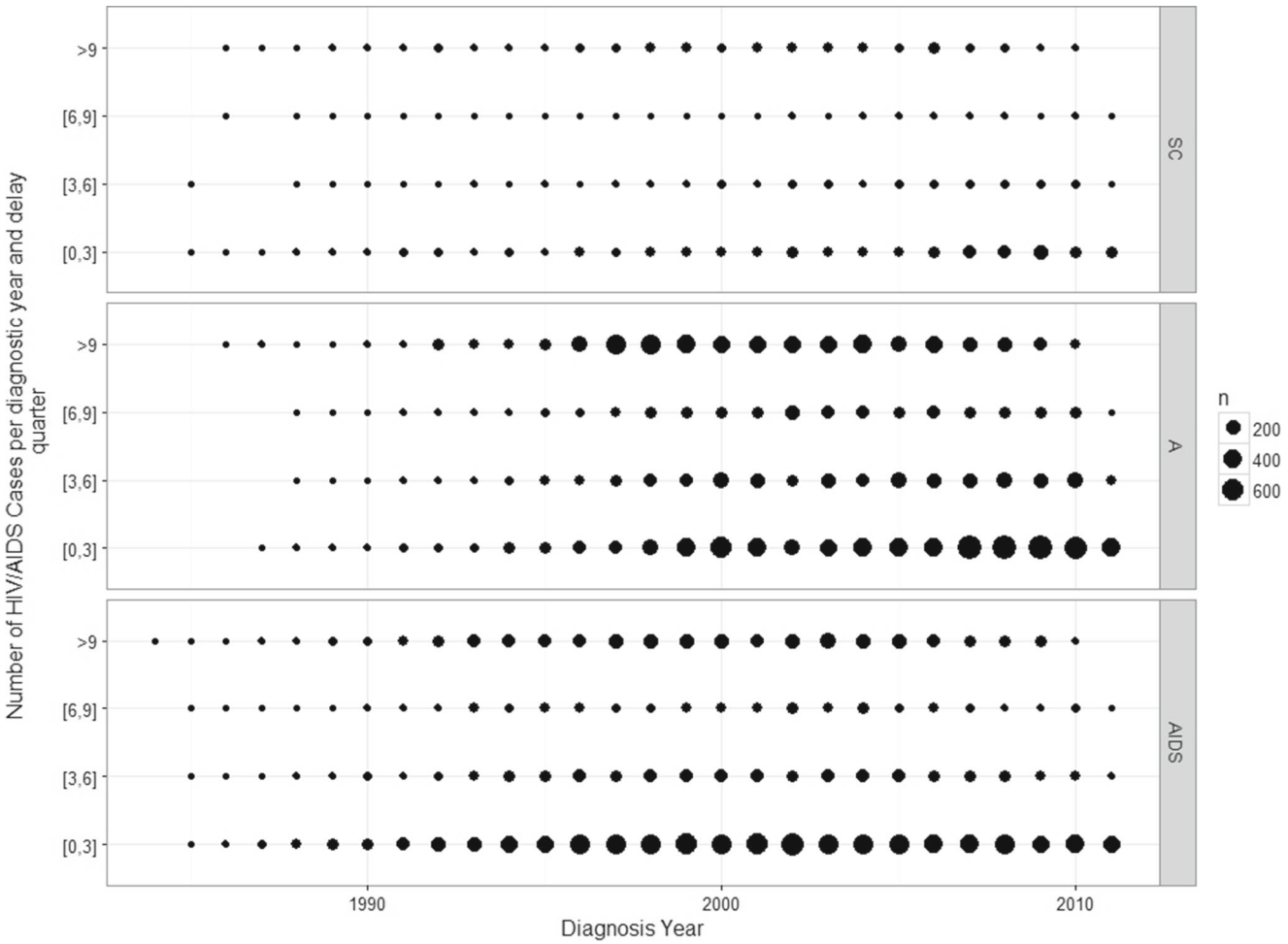

Fig. 6 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and disease status (*right vertical boxes*)

models [32]. In particular feedfoward multilayer perceptron networks with back-propagation training algorithms are the most widely used. They are effective in the analysis of complex data with non-linear trends and time-dependent covariates, and even high-order interactions [37].

In this paper we implemented a feed-forward multilayer perceptron with a hidden layer (with 18 neurons). Sigmoid functions were used as transfer and activations functions. For training, a back propagation algorithm with 500 training cycles, a learning rate of 0.3, a momentum of 0.2, and an error of $\epsilon = 1.0E^{-5}$ was used. The constants were finely tuned according to the data and results.

Naive bayesian classifier A naive (or simple) Bayesian (NB) classifier is a probabilistic classifier which assumes that all attributes contribute equally, and independently, to the final decision [21]. It is a computational simple algorithm that can handle a data set with many attributes and thus widely-used in medical data mining.

Support Vector Machines Learning Support Vector Machines (SVM) are amongst the most popular and efficient classification and regression methods currently available. These algorithms apply simple linear methods in a high-dimensional feature space that is non-linearly related to the input space. Usually, all attributes are employed and non-overlapping partitions are generated. In this paper we used a SVM for classification purposes, with a sigmoid kernel of degree 3, a gamma parameter equal to the inverse of sample size, a constant of the regularization term in the Lagrange formulation equal to one, a tolerance equal to 0.001, an error of $\epsilon = 0.1$ and a heuristic shrinking. The training set was randomly chosen and contained 80 % of the all data. The constants were finely tuned according to the data and results.

K - Nearest Neighbor The K-nearest neighbor algorithm is one of the most popular classification algorithm methods. The algorithm performs a case partition in a pre-user-

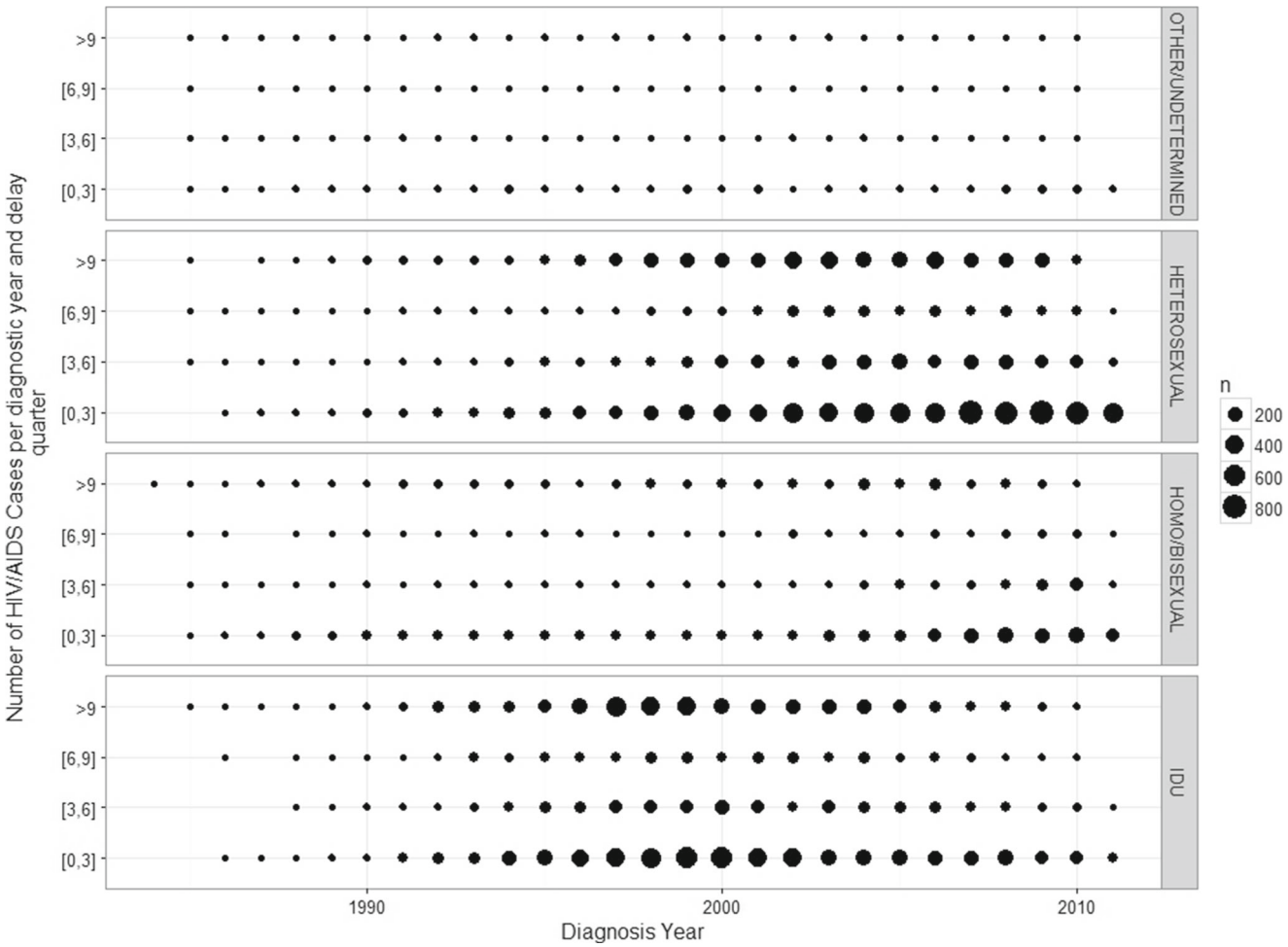

Fig. 7 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and risk group (*right vertical boxes*)

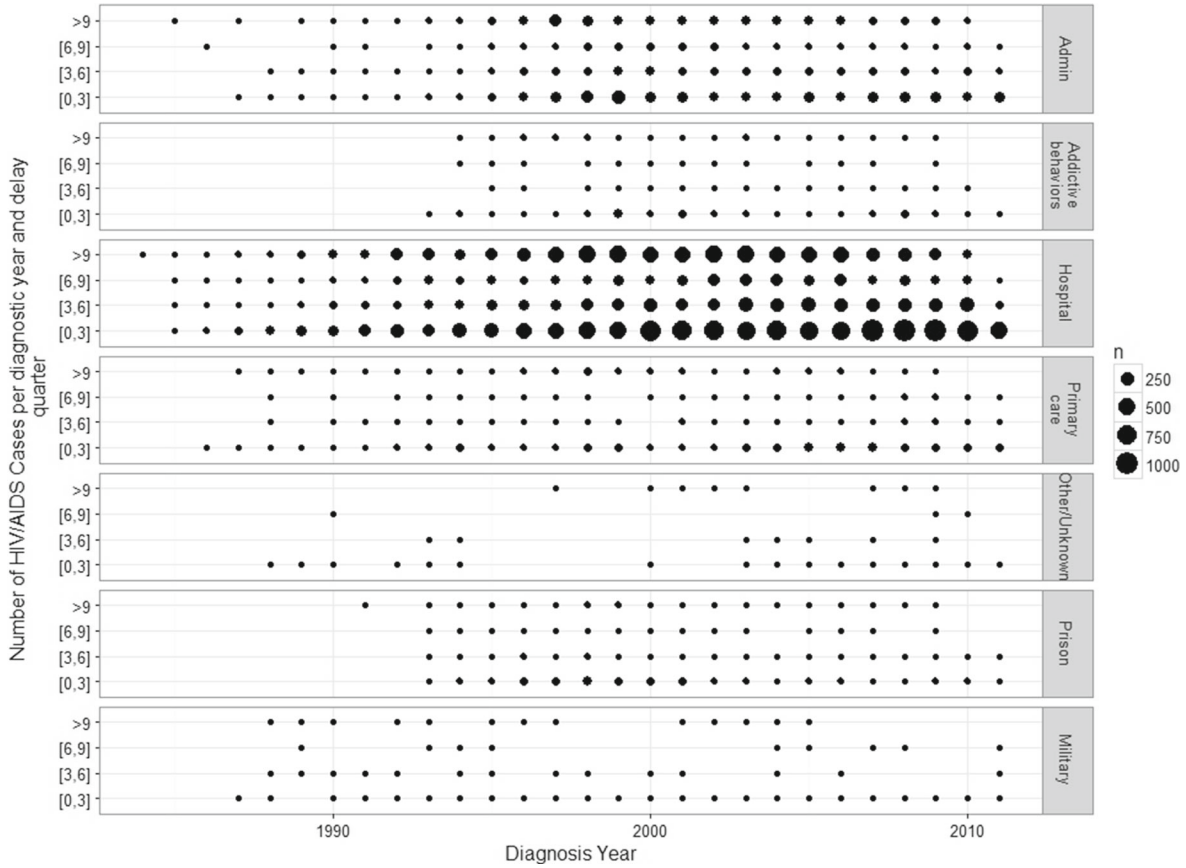

Fig. 8 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and health care institution (*right vertical boxes*)

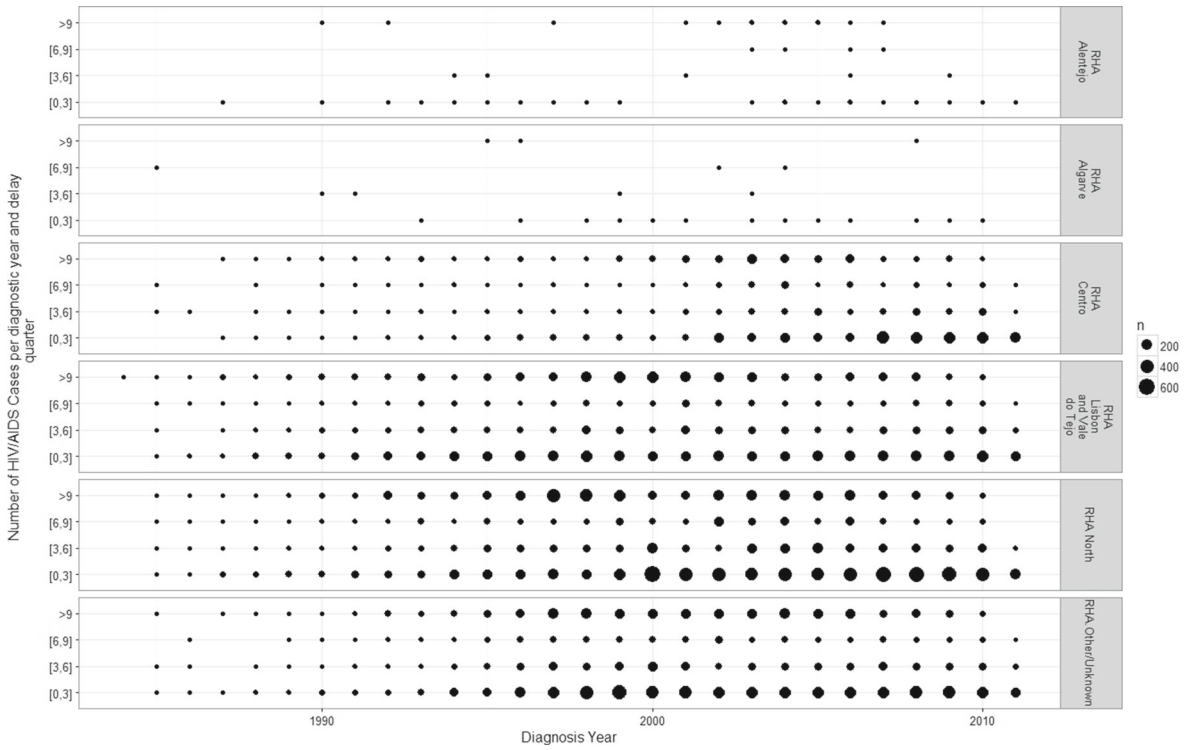

Fig. 9 Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and regional administration (*right vertical boxes*)

Table 1 Two reporting delays classes by input features

		0–3		>3	
		mean	std	mean	std
Age at diagnosis		37.74	13.03	36.21	12.93
		n	%	n	%
Gender	Female	4397	24	4832	26
	Male	13815	76	13466	74
Patient Nationality	Africa	1786	41	1599	53
	Americas	448	2	361	2
	Europe	15309	84	15688	86
classified by HIV-AIDS WHO regions	Low Incidence Countries	30	0	23	0
	Nat.Other - Unknown	639	4	627	3
Disease stage	A	7501	41	9777	53
	AIDS	9063	50	6712	37
	SC	1648	9	1809	10
HIV Risk-group	Heterosexual	8491	47	7683	42
	Homo - Bisexual	2778	15	2575	14
	IDU	6331	35	7452	41
	Other - Undetermined	612	3	588	3
Health care institution	Centres for treatment of addictive behaviours	457	3	296	2
	Unspecific Administration	2346	13	2898	16
	Hospital	13653	75	13913	76
	Primary care	1033	6	716	4
	Prisons	555	3	358	2
	Military	103	1	80	0
	Other - Unknown	65	0	37	0
Regional Health Administration	RHA Alentejo	91	0	109	1
	RHA Alagarve	28	0	10	0
	RHA Centro	2367	13	2222	12
	RHA Lisboa e Vale do Tejo	4103	23	4742	26
	RHA Norte	6581	36	6381	35
	RHA Other - Unknown	5041	28	4834	26

defined number of clusters by comparing a given test sample with a training sample. Each object is assigned to the class corresponding to the majority vote from its K nearest neighbors. "Closeness" is defined in terms of a distance metric. For the given data, the K-means algorithm with 2 clusters was applied, with a mixed euclidean distance given the nominal and the quantitative nature of the input features.

A 10-fold cross-validation was used for MLP, NB and KNN validation. A leave-one-out method was applied to the SVM model.

Table 2 MLP performance

	Confusion matrix				Accuracy	Precision	Recall	Time
	Actual: 0–3	Actual: >3	sum	%				
Predicted: 0–3	13831	9136	22967	60 %	62.98 % +/− 0.78 %	76 %	60 %	50min37s
Predicted: >3	4381	9162	13543	68 %				
%	76 %	50 %						
sum	18212	18298						

Table 3 KNN performance

	Confusion matrix				Accuracy	Precision	Recall	Time
	Actual: 0–3	Actual: >3	sum	%				
Predicted: 0–3	8344	4261	12605	66 %	61.30 % +/- 0.88 %	45 %	66 %	2 min
Predicted: >3	9868	14037	23905	59 %				
%	45 %	77 %						
sum	18212	18298						

Results

The observed number of diagnosed HIV-AIDS cases (in all stages) in Portugal, from 1983 to 2011, is presented in Fig. 1. These numbers have to be interpreted according to the above mentioned surveillance system changes. The number of HIV-AIDS cases increased between 1983 and 2000, four years after the introduction of HAART. When the notification became mandatory, a slight growth was observed 1. The prevalence of HIV-AIDS changes slightly after these moments.

Figure 2 depicts the reporting delays that are registered in the national HIV-AIDS surveillance system. The delays were grouped into trimesters and the annual percentage of cases within each diagnosis year is represented. The most recent year (shaded region) does not seem to be describing the real situation as several cases have not been notified yet.

Most of the cases are reported within 3 months after diagnosis but some are still being reported with more than one year of delay. For the sake of clarity, and since they exist in low numbers, delays longer than 18 months are omitted from Fig. 2.

In Fig. 3 it can be seen that the majority of cases were reported in most recent years of diagnosis and with a delay of 0-3 months.

The longitudinal distributions of the reporting delays according to several variables (Figs. 4, 5, 6, 7, 8 and 9) pictures the most descriptive features of the infection in Portugal. The infection is positively associated with the male gender and most of the cases have naturally a European nationality, followed by Africans.

Early detections, corresponding to asymptomatic cases, corresponded to most of the cases, especially in the most recent years, and presented equally distributed reporting delays through the quarters. Large concentrations of AIDS-cases (late detections and / or disease progressions) must also be noticed, mostly with a reporting delay lower than 3 months.

As for the risk-groups, the infection has been most prevalent amongst the heterosexual community, mainly after 1999, but a large frequency of IDUs is also visible from 1996 to 2003.The cases are equally distributed through the different reporting delays and diagnosis years.

The major reporting institutions are the hospitals, and most of their reporting delays are lower than 3 months.

Considering the Regional Health Administration it can be seen that RHA Norte had the majority of the reported cases. It can also be seen that the time lag between diagnosis and reporting is, in most cases, less than 3 months.

The distribution of the binary variable for the reporting delays according to sex, age, nationality, disease stage, HIV risk-group, type of Health care institution and Regional Health Administration, is balanced for the two classes, Table 1 .

Tables 2, 3, 4 and 5 describe the performance of the data mining algorithms. The accuracy in predicting the class membership ranged from 53 % (NB) to 63 % (MLP) approximately, the precision ranged from 16 % (NB) to 76 % (MLP) and finally the recall ranged from 60 % (MLP, NB and SVM) to 66 % (KNN). The fastest algorithm needed 15s to produce the results and the slowest approximately 51min.

Table 4 NAIVE performance

	Confusion matrix				Accuracy	Precision	Recall	Time
	Actual: 0–3	Actual: >3	sum	%				
Predicted: 0–3	2997	1983	4980	60 %	52.90 % +/- 0.97 %	16 %	60 %	15s
Predicted: >3	15215	16315	31530	48 %				
%	16 %	11 %						
sum	18212	18298						

Table 5 SVM performance

	Confusion matrix				Accuracy	Precision	Recall	Time
	Actual: 0–3	Actual: >3	sum	%				
Predicted: 0–3	12795	8412	21207	60 %	62 %	70 %	60 %	2min
Predicted: >3	5417	9886	15303	65 %				
%	70 %	54 %						
sum	18212	18298						

Conclusions

Surveillance systems rely on processes using pre-specified diseases case definitions and employ manual data collection, human decision making, and manual data entry. The analysis of incidence and prevalence should thus include the analysis of the historical events that may affect the way the systems collect the data.

Considering the reporting delay divided in quarters a 2-class division arises naturally, with a cut-off at the 3-month delay. According to this classification, several supervised learning techniques were applied.

Analyzing the behavior of these two groups according to the year of diagnosis, it can be seen that it is mostly constant until the most recent years (in the last years the percentage of cases is biased due to the reporting delay). Moreover, the two groups seemed to behave similarly with respect to the patient's age at the diagnosis, gender, stage of the disease, transmission risk group, nationality, type of the health care provider that made the diagnosis and administrative and financial responsible from the health care provider.

Tables 2, 3, 4 and 5 show that MLP provided the best results, with a higher classification accuracy (approximatelly 63 %), precision (approximatly 76 %) and recall (aproximatly 60 %), on the other hand it is considerably slower. It can predict, with reasonable efficiency the group of reporting delays less than 3 months long. The SVM model has similar results and is considerable faster.

While, around 60 % of accuracy may be a reasonable result it can be explained by characteristics of the input data. In many cases,the quality of the data within the biomedical and healthcare fields is inferior to that found in other fields [21]. In our data set the main reasons for the poor classification quality are most probably related to stigma around the disease that leads patients to provide incorrect informations (mainly in transmission group), to high demands of the healthcare systems and to implementation of the surveillance system, more specifically paper form reports and poor communication between the stakeholders. In a previous qualitative assessment of the Portuguese Surveillance System, Mauch pointed out that all clinicians reported that they complete the notification form after the patient has left the office, sometimes several days or weeks later. This practice has the potential to contribute to inaccuracies in reporting for some variables, such as the associated risk group, due to recall errors[14]. Moreover, errors may also arise from the transcription of the information in the paper report to databases.

Another important issue is that simply there is no connection between the measured features and the reporting delay classes.

Acknowledgements A. Rita Gaio was partially supported by CMUP (UID/MAT/00144/2013), which is funded by FCT (Portugal) with national (MEC) and European structural funds (FEDER), under the partnership agreement PT2020. Luís Paulo Reis was partially by the European Regional Development Fund through the programme COMPETE by FCT (Portugal) in the scope of the project PEst - UID/CEC/00027/2015 Luís Paulo Reis and Brígida Mónica Faria were partially funded by QVida+: Estimação Contínua de Qualidade de Vida para Auxílio Eficaz à Decisão Clínica, NORTE-01-0247-FEDER-003446, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement.

References

1. J. U. N. P on HIV-AIDS et al.: A framework for monitoring and evaluating HIV prevention programmes for most-at-risk populations Geneva. Joint United Nations Programme on HIV-AIDS, 2008.
2. J. U. N. P on HIV-AIDS, and World Helath Organization: Guidelines on surveillance among populations most-at-risk for HIV Geneva. Joint United Nations Programme on HIV-AIDS, 2011.
3. Hutchinson, J., The biology and evolution of HIV. *Annu. Rev. Anthropol. JSTOR*, 85–108, 2001.
4. Centers for disease control and prevention, HIV transmission, 2015.
5. WHO, HIV AIDS: Fact sheet N. 360, World Health Organization, 2014.
6. WHO: Health in 2015: from MDGs to SDGs World Health Organization, World Health Organization, 2015.
7. Barnard, J., and Meng, X. L., Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat. Methods Med. Res., SAGE Publications* 8:17–36, 1999.
8. Brookmeyer, R., and Gail, M. *AIDS epidemiology: a quantitative approach*. Oxford: Oxford University Press, 1994.
9. Teutsch, S., and Churchill, R. *Principles and practice of public health surveillance*. Oxford: Oxford University Press, 2000.

10. Stoto, M. A.: Public health surveillance: a historical review with a focus on HIV AIDS. RAND Corporation, 2003.
11. European Center for Disease Prevention and Control, Data quality monitoring and surveillance system evaluation – A handbook of methods and applications. Stockholm: ECDC, 2014.
12. Waller, L. A., and Gotway, C. A. *Applied spatial statistics for public health data*. Vol. 368. New York: Wiley, 2004.
13. European Center for Disease Prevention and Control, HIV AIDS surveillance in Europe 2011, Stockholm: European Centre for Disease Prevention and Control, 2012.
14. Mauch, S.: Situational assessment of the HIV AIDS notification system - a Portuguese experience national coordination for HIV infection, 2009.
15. Bouman, P., Dukic, V., and Meng, X.-L., A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Stat. Sin., JSTOR*, 325–357, 2005.
16. Cui, J., and Kaldor, J., Changing pattern of delays in reporting AIDS diagnoses in Australia. *Aust. N. Z. J. Public Health, Wiley Online Library* 22:432–435, 1998.
17. Green, T. A., Using surveillance data to monitor trends in the AIDS epidemic. *Stat. Med., Wiley Online Library* 17:143–154, 1998.
18. Harris, J. E., Reporting delays and the incidence of AIDS. *J. Am. Stat. Assoc., Taylor and Francis Group* 85:915–924, 1990.
19. Kalbfleisch, J., and Lawless, J., Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Stat. Sin., World Scientific Publishing* 1:19–32, 1991.
20. Jacob, S. G., and Ramani, R. G., Data mining in clinical data sets: a review training. *Citeseer*, 4, 2012.
21. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., and Hua, L., Data mining in healthcare and biomedicine: a survey of the literature. *J. Med. Syst., Springer* 36:2431–2448, 2012.
22. Dunham, M.: Data mining: introductory and advanced topics Pearson education, 2006.
23. Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., and Geissbuhler, A., Clinical data mining: a review. *Yearb. Med. Inform.* 121–133, 2009.
24. Rosenberg, P., A simple correction of AIDS surveillance data for reporting delays. *J. Acquir. Immune Defic. Syndr.* 3:49–54, 1990.
25. Brookmeyer, R., and Liao, J., Statistical modelling of the AIDS epidemic for forecasting health care needs. *Biometrics, JSTOR*, 1151–1163, 1990.
26. Brookmeyer, R., and Damiano, A., Statistical methods for short-term projections of AIDS incidence. *Stat. Med., Wiley Online Library* 8:23–34, 1989.
27. Brookmeyer, R., and Gail, M. H., A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *J. Am. Stat. Assoc., Taylor and Francis Group* 83:301, 1988.
28. Pagano, M., Tu, X. M., De Gruttola, V., and MaWhinney, S., Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. *Biometrics, JSTOR*, 1203–1214, 1994.
29. Harris, J. E. *Delay in reporting acquired immune deficiency syndrome (AIDS)*. Cambridge, Mass., USA: National Bureau of Economic Research, 1987.
30. Midthune, D. N., Fay, M. P., Clegg, L. X., and Feuer, E. J., Modeling reporting delays and reporting corrections in cancer registry data. *J. Am. Stat. Assoc., Taylor and Francis* 100:61–70, 2005.
31. Wasserstein, R. L., and Lazar, N. A.: The ASA's statement on p-values: context, process, and purpose. *Am. Stat.*, 2016.
32. Sharaf, T., and Tsokos, C. P., Two artificial neural networks for modeling discrete survival time of censored data. *Advances in Artificial Intelligence, Hindawi Publishing Corp.* 2015:1, 2015.
33. European Centre for Disease Prevention and Control WHO Regional Office for Europe, Hiv aids surveillance in Europe. Tech. Report, Stockholm, European Centre for Disease Prevention and Control, 2008.
34. Portaria n. 258 2005 de 16 de março. diário da repúbicla n. 53 2005 - i série b. Ministério da saúde. Lisboa.
35. Oliveira, A., Costa, J., and Gaio, A. R., The incidence of AIDS in Portugal adjusted for reporting delay and underreporting. In: 9th Iberian conference on information systems and technologies (CISTI), pp. 1–5: IEEE, 2014.
36. Amaral, J., Pereira, E., and Paixão, M., Data and projections of HIV AIDS cases in Portugal: an unstoppable epidemic? *J. Appl. Stat., Taylor and Francis* 32:127–140, 2005.
37. Yang, Y.: Neural Network Survival Analysis, Faculty of Sciences - Department of Applied mathematics and computer science, 2011.
38. Barros, P., Machado, S., and Simões, J., Portugal: health system review. *Health Syst. Transit.* 13(4):1–156, 2011.