MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190224

What Is a Chronic Disease? A Contribution Based on the Secondary Use of 161 Million Discharge Records

Emeric Mellot[a], Thibaut Balcaen[b], Matthieu Calafiore[c], Guillaume Bouzillé[d], Jean-Baptiste Beuscart[e], Grégoire Ficheur[e], Emmanuel Chazard[e]

[a] Univ. Lille, CERIM EA2694, General practitioner, F-59000 Lille, France
[b] Univ. Lille, CERIM EA2694, F-59000 Lille, CH Saint Quentin, F-02100 Saint Quentin, France
[c] Univ. Lille, CERIM EA2694, F-59000 Lille, General practitioner, F-59150 Wattrelos, France
[d] Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France
[e] Univ. Lille, CHU Lille, CERIM EA2694, F-59000 Lille, France

Abstract

Several definitions of chronic diseases exist. The objective is to reuse a nationwide medical-administrative database (PMSI) to estimate the lifespan of diagnostic codes, hence the chronicity of the corresponding diseases. We analyzed 162 million inpatient stays from 2008 to 2014, and estimate the lifespan of every ICD-10 code for every patient, identified by a unique imprint. We calculated 200 indicators for different time and survival values, and selected the ones that maximized the area under the ROC curve (AUC) drawn by comparison against 4 chronic disease classifications: CCI, ALD, result from the analysis of ICD-10 labels, and a handmade list. The best indicator was the time to reach a survival of 4.5%. It enables to get the following AUC: 78.9% compared with CCI, 90.3% compared with ALD, 75.1% compared with labels analysis, and 91.5% compared with the handmade list. This indicator enables to classify 23,349 ICD-10 codes from "most chronic" to "most acute". The 100 most chronic codes are listed.

Keywords:
Chronic disease, Patient discharge, Big data.

Introduction

Chronic conditions are the world epidemic of the 21st century, accounting for about two-third of all deaths [1–4]. For the WHO, a chronic condition is of long duration and generally slow progression [5]. Other definitions propose a chronic/non-chronic classification according to time thresholds which vary from 3 to 12 months [2,6]. In France, the compulsory health insurance provides with a list of 30 categories of chronic conditions, called "ALD List" [7]. In the USA, the American Agency for Healthcare Research and Quality publishes the Chronic Condition Indicator (CCI) which provides a chronic/acute status for all ICD-10-CM codes [8].

The French nationwide hospital discharge database (PMSI) comprehends all the impatient stays from both profit and non-profit hospitals. This exhaustive nationwide database comprehends about 25 to 30 million inpatients stays per year [9]. For each inpatient stay, the data notably comprehend the ICD-10 principal and secondary diagnostic codes, the anonymized date of the stay, and a unique cryptographic patient identifier. This identifier enables to follow the readmissions of any patient without disclosing his identity, even in different hospitals. The PMSI data are initially collected for billing purposes.

Our objective is to reuse the big data [10,11] from the PMSI to estimate the temporal persistence of the ICD-10 codes of French patients, in order to propose a ranking of the ICD-10 codes according to their chronicity.

Methods

Population of interest

We processed the anonymous data from the French PMSI from 2008 to 2014, which comprehends 162 million stays after exclusion of sequential treatments (e.g. hemodialysis, radiotherapy, etc.). An authorization was issued from the French agency for data protection (CNIL) [9].

Computation step

The stays were sorted by inpatients identifiers thanks to their unique imprints. Patients who wished to remain anonymous (e.g. for an abortion) accounted for 0.4% of all stays and were excluded. The sample then comprehended 46 million patients, 161 million inpatient stays, and 503 million ICD-10 codes.

For each inpatient, we listed all the codes found during the observation period (2008-2014), then we looked when they appeared and disappeared, to calculate their lifespans. We assumed the date of every diagnostic code was the first day of the inpatient stay. Then, if a code was first found in a stay at date *t1* and last found in a stay at date *t2*, its lifespan was *(t2 – t1 + 1)*. In order to simplify computations, we did not implement right-censoring. A code "disappeared" when it was not followed by itself (or any code having the same 3 first digits) within 720 days (this method was determined after several tests, using a 10% training sample). For instance, "J18.1" and "J18.9" both relate to "J18". For a given patient, the presence of "J18.9" will extend the lifespan of "J18.1". From our experience, those codes variations are more often due to inaccurate coding than patient's disease evolution.

This way, for every ICD-10 code and every patient, we could compute a duration. Those durations were plotted through survival curves which, in the absence of censoring, correspond to empirical distribution functions of the durations [12]. From those curves, we could compute the survival at a given time

(we tested 100 times, from 1 to 100 days), or the time to get a certain survival (we tested 100 survival rates, from 50% to 0,5%). For each ICD-10 code, we could then store 200 values.

We used the R programming language with the additional packages data.table, survival, and proc [13].

Validation step

Overview

The external validation was performed against 4 lists of diseases, presented hereafter:

- "**CCI**": list of chronic ICD-10-CM codes provided by the American Agency for Healthcare Research and Quality published in 2017 [4,8], and mapped to French ICD10 codes.
- "**ALD**": list of 30 "long duration diseases" provided by the French compulsory insurance in 2015, and mapped to 122 3-digit ICD-10 codes [7]
- "**ICDlabel**": the result from text analysis of the wordings of all French ICD-10 codes [14,15]
- "**Handmade100**": a list of 100 chronic diseases identified from the ICD-10 from our knowledge.

Preparation of the "CCI" list

The 2017 American CCI provides the chronicity of 64,955 ICD-10-CM codes [4,8] while the 2017 French ICD-10-FR has 40,519 codes. A mapping between both classifications was realized as follows. For 6,003 codes, a direct matching was possible. Then, when a French code did not have any match, the ICD-10-CM codes starting with the same characters were considered. For a given ICD-10-FR code, if all the corresponding codes were chronic or non-chronic in the CCI classification, the French code could be classified as chronic or non-chronic. If the corresponding codes were of mixed type, the French code couldn't be classified. We retrieved 11,034 codes, 10,487 of which were found in our population. We kept the 4,323 chronic codes and selected the 4,323 most frequent non-chronic codes to obtain a total list of 8,626 codes.

Preparation of the "ALD" list

Among the 23,349 codes of our database, 1,733 were matching the 122 3-digits codes from the official 2015 ALD list [7]. We made the simple assertion that all the conditions which were not in the ALD list were non-chronic.

Preparation of the "ICDlabel" list

We searched for the keywords "acute" and "chronic" in the free-text labels of the 2017 French version of ICD-10 [14,15]. We removed all the results which contained terms like "sub-acute", "not specified as acute or chronic" or "antecedent of acute/chronic", etc. We then obtained 283 "acute" codes and 236 "chronic" codes. Among acute codes, we only kept the 236 most frequent ones.

Preparation of the "Handmade100" list

For each ICD-10 code of the database, we calculated how many patients had this code at least once, in order to get the cumulated prevalence of those codes during the period. We sorted the codes by decreasing prevalence, reviewed them, and selected the codes for which we fully agreed with the CCI classification, until we got 100 chronic and 100 acute codes.

Indicators selection

For this step, the statistical individual was the patient. The aim was to discriminate "chronic" codes from "acute" codes (other codes were ignored). For each of the 200 available indicators, we draw a ROC curve [16] against each of the 4 validation lists and calculated the AUC. For each of the 4 validation lists, we then selected the indicator that obtained the highest AUC. With the best indicator, we then selected the threshold giving the closest values of sensibility and specificity (respectively *Se* and *Sp*, which are only provided for descriptive purposes). In some cases, a sensibility analysis was also performed, as described hereafter.

Results

Example of survival curve

For a typical chronic condition, such as end-stage renal disease (Figure 1), the survival curve begins with a moderate drop, then decreases slowly. The drop corresponds to the patients in which either the code is present once, or our algorithm failed to detect any recurrence of the code. The survival at the end of our study is always zero as we did not implement right-censoring.

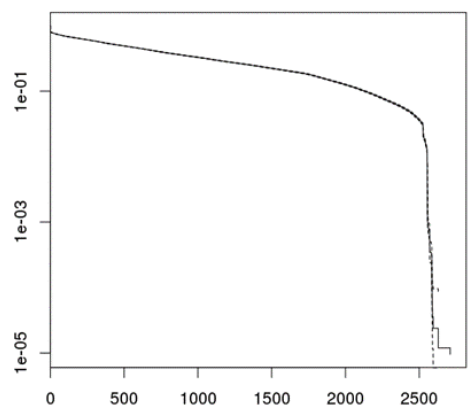

Figure 1. Survival curve of a chronic code, N180 (end-stage renal disease; time in days; Y axis in logarithmic scale)

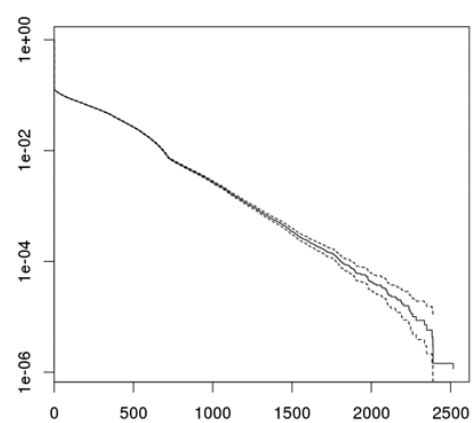

Figure 2. Survival curve of an acute code, E559 (vitamin D deficiency; time in days; Y axis in logarithmic scale)

For a typical non-chronic condition, such as vitamin D deficiency (Figure 2), the initial drop is important, and the survival rate rapidly approaches zero.

It is worth noting that these curves do not show the actual duration of the diseases, but only enable to estimate their chronicity.

Indicator selection

Overview

In this step, for each validation list, and for each of the 200 indicators, we draw a ROC curve and selected the indicator associated with the highest AUC value.

Indicator selection for "CCI"

Compared with the "CCI" list (4,323 chronic codes, and 4,323 acute codes), the best indicator is the time required to reach a survival of 4.5%. It obtains an AUC of 0.789 [0.780; 0.799] (Figure 3). The best point, with a threshold of 182 days, gets a specificity of 72.4% and a sensitivity of 72.7%.

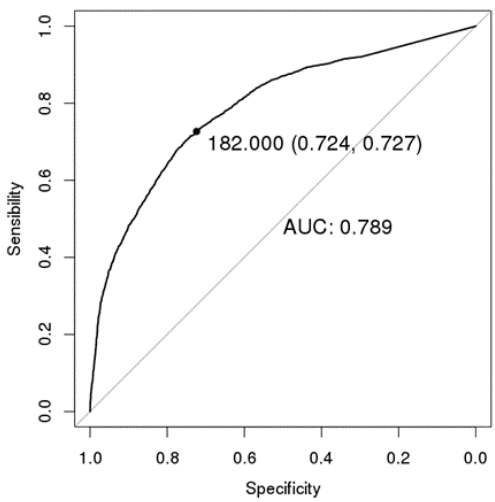

Figure 3. ROC curve of the indicator "time until survival=4.5%" against the "CCI" validation list

A first sensibility analysis, restricting the evaluation to the 6,003 codes (over 8,646) having an exact match between ICD-10-FR and ICD-10-CM, enables to select the same indicator too, with AUC=0.8, Sp=73.2% and Se=73.3% for a threshold of 218 days.

In a second sensibility analysis, we only included the codes that were used in only *n* patients, and reran the analysis. With the threshold n≥10, the AUC rises up to 0.817 (with the same indicator). For a threshold n≥1000, the AUC rises up to 0.83 (with another indicator, the survival rate at 85 days).

The third sensibility analysis consisted of including all the 10,487 codes from the CCI data. We then obtain an AUC of 0.796 with the same indicator, and Sp=73.2% and Se=73.4% for a threshold of 174 days.

Indicator selection for "ALD"

Compared with the "ALD" list (1,733 chronic codes, and 21,616 non-chronic codes), the best indicator is the time required to reach a survival of 7%. It enables to get an AUC of 0.903 [0.897; 0.909] (Figure 4). The best point, with a threshold of 168 days, gives Sp=Se=82.7%.

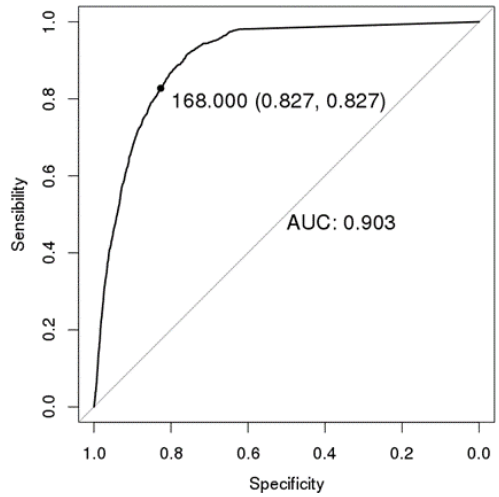

Figure 4. ROC curve of the indicator "time until survival=7%" against the "ALD" validation list

The sensibility analysis consisted of choosing manually the same indicator as previously, namely the time required to reach a survival of 4.5% (instead of 7%). This indicator obtains an AUC of 0.899, and the best threshold (322 days) enables to reach Sp=Se=82.2%.

Indicator selection for "ICDlabel"

Compared with the "ICDlabel" list, the best indicator is the time required to each a survival of 4.5%. It obtains an AUC of 0.751 [0.707; 0.794]. The best point, with a threshold of 210 days, gives Sp=Se=61.9%.

The sensibility analysis consists of comparing this list of codes to the "CCI" list. The comparison between both lists leads to a specificity of 81.1% and a sensibility of 85.5%. By removing the codes from the "ICDlabel" list which are not in the "CCI" list, we obtain an AUC of 0.774 for the same indicator, and the best threshold is 200 days, with Se=70.6% and Sp=70.1%.

Indicator selection for "Handmade100"

Compared with the "Handmade100" list (100 chronic codes, and 100 acute codes), the best indicator is the time required to reach a survival of 4.5%. It obtains an AUC of 0.915 [0.876; 0.954]. The best point, with a threshold of 258 days, enables to reach Sp=Se=84%.

The first sensibility analysis consisted of refactoring this list by exact matching with the CCI list, and not by expert-operated construction, for both the chronic codes list, and the acute codes list. We then obtain an AUC of 0.81 with the same indicator, and the threshold of 317 days enables to reach Sp=Se=74%.

The second sensibility analysis did the same but with the 100 less frequent chronic and acute codes. We then obtain AUC=0.911 for the same indicator, and the threshold of 102 days leads to Sp=85% and Se=86%.

List of "chronic" ICD-10 codes

The four validation steps enabled to identify that, among the 200 indicators, the indicator "time until survival=4.5%" was the most accurate, as it was nearly always selected regarding the AUC criteria. The ICD-10 codes were sorted by decreasing value of this indicator, which enabled to select the 100 "most chronic" conditions, which are presented in Table 1.

Table 1. List of top-100 chronic ICD-10 codes

#	Code	Wording
1	Q861	Fetal hydantoin syndrome
2	E761	Mucopolysaccharidosis, type II
3	D595	Paroxysmal nocturnal hemoglobinuria
4	M834	Aluminum bone disease
5	E763	Mucopolysaccharidosis, unspecified
6	D800	Hereditary hypogammaglobulinemia
7	E760	Mucopolysaccharidosis, type I
8	M0511	Rheumatoid lung disease with rheumatoid arthritis of shoulder
9	E849	Cystic fibrosis, unspecified
10	Z941	Heart transplant status
11	Z942	Lung transplant status
12	E840	Cystic fibrosis with pulmonary manifestations
13	E848	Cystic fibrosis with other manifestations
14	N180	End stage renal disease
15	G243	Spasmodic torticollis
16	Z943	Heart and lungs transplant status
17	G245	Blepharospasm
18	G513	Clonic hemifacial spasm
19	G241	Genetic torsion dystonia
20	G242	Idiopathic nonfamilial dystonia
21	E841	Cystic fibrosis with intestinal manifestations
22	Z948	Other transplanted organ and tissue status
23	D570	Sickle-cell anemia with crisis
24	Z9481	Bone marrow transplant status
25	E710	Maple-syrup-urine disease
26	E711	Other disorders of branched-chain amino-acid metabolism
27	G244	Idiopathic orofacial dystonia
28	Q812	Epidermolysis bullosa dystrophica
29	C83	Non-follicular lymphoma
30	M8336	Adult osteomalacia due to malnutrition, leg
31	N16	Renal tubulo-interstitial disorder in diseases classified elsewhere
32	D86	Sarcoidosis
33	Z944	Liver transplant status
34	E762	Other mucopolysaccharidosis
35	B676	Echinococcus multilocularis infection, other and multiple sites
36	Z940	Kidney transplant status
37	E752	Other sphingolipidosis
38	D562	Delta-beta thalassemia
39	D571	Sickle-cell disease without crisis
40	M0508	Felty's syndrome, other joints
41	G120	Infantile spinal muscular atrophy, type I [Werdnig-Hoffman]
42	M0720	Psoriasic spondylopathy, multiple sites
43	E740	Glycogen storage disease
44	M05	Rheumatoid arthritis with rheumatoid factor
45	E21	Hyperparathyroidism and other disorders of parathyroid gland
46	M349	Systemic sclerosis, unspecified
47	M340	Progressive systemic sclerosis
48	Z491	Renal dialysis
49	N18	Chronic kidney disease (CKD)
50	M0758	Enteropathic arthropathies, other joint
51	W443	Foreign body, eye
52	E661	Drug-induced obesity
53	M0503	Felty's syndrome, wrist
54	D839	Common variable immunodeficiency, unspec.
55	M058	Other rheumatoid arthritis with rheumatoid factor
56	M341	CR(E)ST syndrome
57	E662	Morbid (severe) obesity with alveolar hypoventilation
58	M0801	Unspecified juvenile rheumatoid arthritis, shoulder
59	B24+0	Pre-AIDS
60	D812	Severe combined immunodeficiency with low or normal B-cell numbers
61	M9411	Relapsing polychondritis, scapular region
62	B675	Echinococcus multilocularis infection of liver
63	M0833	Juvenile rheumatoid arthritis (seroneg.), wrist
64	Q819	Epidermolysis bullosa, unspecified
65	E220	Acromegaly and pituitary gigantism
66	Q442	Atresia of bile ducts
67	D817	Major histocompatibility complex class II deficiency
68	M0505	Felty's syndrome, hip
69	K51	Ulcerative colitis
70	G710	Muscular dystrophy
71	G121	Other inherited spinal muscular atrophy
72	L103	Brazilian pemphigus [fogo selvagem]
73	M348	Other forms of systemic sclerosis
74	E723	Disorders of lysine and hydroxylysine metabolism
75	M339	Dermatopolymyositis, unspecified
76	E771	Defects in glycoprotein degradation
77	E241	Nelson's syndrome
78	K501	Crohn's disease of large intestine
79	M0581	Other rheumatoid arthritis with rheumatoid factor of shoulder
80	E84	Cystic fibrosis
81	M0580	Other rheumatoid arthritis with rheumatoid factor of unspecified site
82	K508	Crohn's disease of both small and large intestine
83	D838	Other common variable immunodeficiencies
84	M059	Rheumatoid arthritis with rheumatoid factor, unspecified
85	K509	Crohn's disease, unspecified
86	Z992	Dependence on renal dialysis
87	D572	Sickle-cell/Hb-C disease
88	M072	Psoriasic spondylopathy
89	M0590	Rheumatoid arthritis with rheumatoid factor, unspecified, multiple site
90	B677	Echinococcus multilocularis infection, unspec.
91	G248	Other dystonia
92	Z949	Transplanted organ and tissue status, unspec.
93	Z9488	Other transplanted organ and tissue status
94	D830	Common variable immunodeficiency with predominant abnormalities of B-cell numbers and function
95	E724	Disorders of ornithine metabolism
96	M330	Juvenile dermatopolymyositis
97	E753	Sphingolipidosis, unspecified
98	M6110	Myositis ossificans progressiva, unspec. site
99	N188	Other chronic kidney disease
100	M0753	Enteropathic arthropathies, wrist

Discussion

In this paper, we proposed an automated method to identify chronic conditions from ICD-10 codes. We validated the results against four different classifications, and obtained good and stable results. We finally proposed a list of 100 "most chronic" ICD-10 codes.

French rules specify that a diagnostic code should be used when the disease has had a significant impact on the medical management of the patient, which means not only the disease was still active, but also the disease had to be actively treated during the inpatient stay. According to that definition, our work should not enable to identify chronic conditions, but conditions which induce a chronic need for care.

We provide a ranking of the ICD-10 codes according to their chronicity instead of a binary classification. It is possible to calculate and compare the indicator for different periods or different inpatient populations. Results can be used for many applications without the need for a threshold.

Our work is based on years of nationwide data validated by physicians. This database does not comprehend ICD-9 codes. It does not rely on patient self-report, is not limited in time like a cross-sectional study, is not limited to local morbidity registries and does not depend on the judgment of a group of experts.

As our data relate to inpatient stays, codes for non-severe diagnostics are scarce. There is poor tracking, and stays can be several years apart. The evolution of a disease can be different between outpatients and inpatients, e.g. the reason for the admission can be a surgery that cures the chronic condition. The data may be biased due to per-service pricing, as coders pay more attention to codes that bring money to their hospital.

The CCI classification did not perfectly fit our task. It has additional conditions to define whether a code is a disease or not, thus a code can be chronic without being a chronic condition. Our population consisted of inpatients, while the CCI is determined for all patients. As the CCI is made for the American population, some discrepancies might result from differences between French and American morbidities.

The list of the French ALD was not a very good control list either, as it only relates to few chronic conditions, which are far better coded in the PMSI than other conditions. It gives an estimate of how our indicator would perform if the coding was perfect.

Conclusion

We designed a quantitative indicator of the chronicity of the ICD-10 diagnostic codes by reusing 161 million stays of the French hospital nationwide discharge database from 2008 to 2014. This indicator has positive and negative likelihood ratios of 2.63 and 0.38 respectively when compared to the CCI from the AHRQ, and of 4.78 and 0.21 respectively when compared to the ALD from the French public health insurance. This indicator enabled to rank 23,349 ICD-10 diagnostic codes according to their chronicity.

References

[1] World Health Organization, Noncommunicable diseases now biggest killers, *WHO*. (2008). https://www.who.int/mediacentre/news/releases/2008/pr14/en/ (accessed November 25, 2018).

[2] G. Anderson, and J. Horvath, The growing burden of chronic disease in America, *Public Health Rep*. **119** (2004) 263–270. doi:10.1016/j.phr.2004.04.005.

[3] P. Piot, A. Caldwell, P. Lamptey, M. Nyrirenda, S. Mehra, K. Cahill, and A. Aerts, Addressing the growing burden of non-communicable disease by leveraging lessons from infectious disease management, *J Glob Health*. **6** (2016) 010304. doi:10.7189/jogh.06.010304.

[4] W. Hwang, W. Weller, H. Ireys, and G. Anderson, Out-of-pocket medical spending for care of chronic conditions, *Health Aff (Millwood)*. **20** (2001) 267–278. doi:10.1377/hlthaff.20.6.267.

[5] World Health Organization, Major Noncommunicable diseases and their risk factors, *WHO*. (2018). http://www.who.int/ncds/introduction/en/ (accessed November 25, 2018).

[6] A. Clarke, What is a chronic disease? The effects of a re-definition in HIV and AIDS, *Soc Sci Med*. **39** (1994) 591–597.

[7] Ameli (French compulsory health insurance), Affection de longue durée (ALD), (2018). https://www.ameli.fr/assure/droits-demarches/maladie-accident-hospitalisation/affection-longue-duree-ald (accessed November 25, 2018).

[8] Healthcare Cost and Utilization Project (HCUP), Beta Chronic Condition Indicator (CCI) for ICD-10-CM, (2018). https://www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp (accessed November 25, 2018).

[9] French Technical Agency for Hospital Information (ATIH), Nationwide data access for healthcare facilities, researchers and public institutions, (2018). https://www.atih.sante.fr/acces-aux-donnees-pour-les-etablissements-de-sante-les-chercheurs-et-les-institutionnels (accessed November 25, 2018).

[10] E. Baro, S. Degoul, R. Beuscart, and E. Chazard, Toward a Literature-Driven Definition of Big Data in Healthcare, *Biomed Res Int*. **2015** (2015). doi:10.1155/2015/639021.

[11] E. Chazard, G. Ficheur, A. Caron, A. Lamer, J. Labreuche, M. Cuggia, M. Genin, G. Bouzille, and A. Duhamel, Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features, *Stud Health Technol Inform*. **255** (2018) 15–19.

[12] E.L. Kaplan, and P. Meier, Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*. **53** (1958) 457–481. doi:10.1080/01621459.1958.10501452.

[13] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2011. http://www.R-project.org/.

[14] French Technical Agency for Hospital Information (ATIH), ICD-10 FR 2017 for PMSI usage, (2017). https://www.atih.sante.fr/cim-10-fr-2017-usage-pmsi (accessed November 25, 2018).

[15] World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 2018. http://www.who.int/classifications/icd/en/ (accessed November 25, 2018).

[16] J. Fan, S. Upadhye, and A. Worster, Understanding receiver operating characteristic (ROC) curves, *CJEM*. **8** (2006) 19–20.

Address for correspondence
Emmanuel Chazard emmanuel.chazard@univ-lille.fr CERIM, medicine school, 1 av Oscar Lambret, 59035 Lille cedex, France.