EXPLORATION OF THE TIMELINESS OF ICD CODES IN ADMINISTRATIVE DATABASES: A NATIONWIDE STUDY

AUTHOR NAMES AND AFFILIATIONS

Marie SIMON

E-mail address: simon.marie86@gmail.com

Affiliation: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France

Bastien RANCE

E-mail address: bastien.rance@aphp.fr

Affiliations: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France (2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

Sandrine KATSAHIAN

E-mail address: sandrine.katsahian@aphp.fr

Affiliations: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France (2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

Karim BOUNEBACHE

E-mail address: said.bounebache@inserm.fr

Affiliation: (3) INSERM, CépiDc (Epidemiology Centre on Medical Causes of Death), US-10, Kremlin-Bicêtre

Grégoire REY

E-mail address: gregoire.rey@inserm.fr

Affiliation: (3) INSERM, CépiDc (Epidemiology Centre on Medical Causes of Death), US-10, Kremlin-Bicêtre

Gilles CHATELLIER

E-mail address: gilles.chatellier@aphp.fr

Affiliation: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France

Antoine NEURAZ

E-mail address: antoine.neuraz@aphp.fr

Affiliations: (4) Département d'informatique médicale, Necker-Enfants Malades Hospital, Assistance Publique des Hôpitaux de Paris (AP-HP), Paris (2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France (4) LIMSI, CNRS, Université Paris Saclay, Orsay

Anita BURGUN

E-mail address: anita.burgun@aphp.fr

Affiliations: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France (2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France (4) Département d'informatique médicale, Necker-Enfants Malades Hospital, Assistance Publique des Hôpitaux de Paris (AP-HP), Paris

Vincent LOOTEN

E-mail address: lootenv@gmail.com

Affiliations: (1) Département d'informatique médicale, Hôpital Européen Georges Pompidou, AP-HP, Paris, France (2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

CORRESPONDING AUTHOR

Vincent Looten

E-mail address: lootenv@gmail.com

Postal Address:

Hôpital Européen Georges Pompidou

Département d'informatique médicale

20 rue Leblanc

75015 Paris,

FRANCE

ABSTRACT

INTRODUCTION: The ICD codes are ubiquitously available in hospital information systems and have been used in a number of areas such as epidemiology, phenotype-genotype association mining, surveillance of the use of drugs and medical devices and health care evaluation. We aimed to analyze the timeliness of the 3-character ICD-10 codes collected in the French national hospital discharge summary database between 2008 and 2017 and we classified the codes according to their evolution.

MATERIAL AND METHODS: We extracted all 3-character ICD-10 codes from all French hospital discharge summaries between 2008 and 2017. For each code and by the month, we computed a relative frequency; we also computed the overall amplitude of the study period. Temporal clustering, according to the SAX representation, was performed to classify the main evolution patterns.

RESULTS: We extracted 238,334,751 encounters corresponding to 56,621,773 distinct patients. 1,006 ICD codes presented a variation of the relative amplitude of frequencies lower than 50%, 510 codes between 50% and 100% and 521 greater than 100%. Out of the 2,037 codes included in the study, we kept the 1,758 for the temporal clustering. Four clusters were identified, including a global increase and a global decrease patterns.

DISCUSSION: The overall results showed a strong instability (i.e. large variation of frequency over time) of the use of ICD codes over time, with an important variation of the relative amplitude of the frequencies. We distinguished between external factors due to changes in billing, organization, policy or regulation and intrinsic factors due to epidemiological phenomenon. The detailed analysis of profiles show that the same cluster can contain profiles influenced by intrinsic epidemiological or external factors or both. Additional knowledge and sources are probably required to determine automatically the origin of the profile.

KEYWORDS: International Classification of Diseases; Data Mining; Data Quality; Timeliness

ABBREVIATIONS

DQ: Data Quality

EHR: Electronic Health Records

ICD: International Statistical Classification of Diseases and Related Health Problems

OHDSI: Observational Health Data Sciences and Informatics

SAX: Symbolic Aggregate Approximation representation

SNDS: *Système National des Données de Santé* (National Health Data System)

EXPLORATION OF THE TIMELINESS OF ICD CODES IN ADMINISTRATIVE DATABASES: A NATIONWIDE STUDY

1 INTRODUCTION

The use of the diagnostic code for "*Malnutrition*" in the French national health administrative databases saw a six-fold increase between 2008 and 2017. Such finding is highly questionable, and researchers could reasonably wonder whether this increase denotes a real phenomenon in the French population or changes in data collection? Health administrative databases are massive repositories of data composed of medical procedures, prescriptions and diagnoses information[1,2] initially collected for administrative purposes. Therefore, the specifics of the data collection may vary depending on other incentives than epidemiology.

Secondary use of health administrative databases for research and public health has become more and more extensive because these databases exhibit many strengths: (i) they are population-based; (ii) data are collected as part of the routine care process, which makes investigations faster and research costs lower; (iii) the follow-up of specific populations can be traced for years or decades; (iv) the presence of several years of data allows for studying changes over time for numerous variables. Among these variables, diagnostic codes expresses as codes from the International Classification of Disease (ICD)provide standardized diagnoses information at local, national and international levels. In European countries, most hospitals leverage codes from the International Classification of Disease (ICD), version 10[3] for billing purposes. The ICD-10 coding system contains over 70,000 disease codes grouped into 22 higher order categories. These codes are ubiquitously available in hospital information systems and can be reused for research purposes[4]. ICD codes have been used in a number of areas such as epidemiology[5,6], phenotype-genotype association mining[7,8], distribution of risk factors and impact on major clinical outcomes[9–11] , surveillance of the use of drugs and medical devices[12,13], health care evaluation and health economic evaluations[14]. In the context of

large-scale collaboration at continental or international levels, like the Observational Health Data Sciences and Informatics (OHDSI) Network[15,16]), ICD data have been largely used to support federated analyses across institutions, e.g., to identify cancer patients[17], and across countries[18]). Over the last years there has been widespread development of medical data repositories, and the research community has put specific emphasis on enhancing the capacity of algorithms to automatically find and use the data, to analyze the data sets, and to mine the data for knowledge discovery[19]. However, there are several bottlenecks to consider when reusing health administrative data and a misinterpretation of the data can have dramatic consequences for research (and even public health decisions) and lead to invalid conclusions.

Regarding hospital data, Agniel *et al.*[20] note that EHRs capture more information than the simple results recorded. For example, the presence of a laboratory test order and the timing of when it was ordered, regardless of any other information about the test result, is indicative of "organizational" aspects such as, for example, the expertise of the clinicians. The presence of some tests may even have a significant association with the odds ratio of death. We hypothesized that, similarly to the data in the EHR, the disease codes in health administrative databases capture more information than the sheer frequency of the diseases in the population: the information carried by ICD codes in health administrative databases is impacted by intrinsic epidemiological factors, such as outbreaks or environmental events, and by external factors, such as financial or health policies. More precisely, we analyzed the whole set of ICD claim data collected in France between 2008 and 2017 to assess whether the temporal context of the variation of ICD codes could provide insight on the respective influence of epidemiological and external factors.

1.1 Related Works

Overtime the importance of epidemiological and external factors might evolve: a query built at a given time for patient selection may not be relevant at a later date, and lead to errors of

interpretation. Mues *et al.*[21] discussed the influence of the Centers for Medicare and Medicaid Services policy on data collection and coding between 2005 and 2012. The authors show that the percentage of inpatient claims with a diagnosis code for coronary atherosclerosis suddenly increased in 2010 because of the expansion of the number of ICD-9/10 diagnosis and procedure code fields on a claim. Moreover, the number of ICD diagnosis codes expanded from 17,000 to 77,000 with the replacement of the ICD-9 by the ICD-10 in 2015. Sáez *et al.*[22] defined timeliness as "the degree of temporal stability of the data." Rey *et al.*[23] proposed a statistical method that detects abrupt changes of the time course of mortality by cause measured with ICD codes. However, to the best of our knowledge, no study has proposed a model-free temporal clustering of ICD codes and a description of the evolution of ICD codes over time.

1.2 Scope and Objective

In this study, we aimed to analyze the timeliness of the 3-character ICD-10 codes collected in the French national hospital discharge summary database between 2008 and 2017. We classified the codes according to their evolution. We discussed the different classes to identify the importance of epidemiological and external effects.

2 MATERIAL AND METHODS

2.1 Study Design

We performed a retrospective study leveraging the French discharge summary database. We searched for the evolution over time of ICD-10 codes. We used the finest degree of granularity available for dates: namely the month.

2.2 Data source

The French National Health Data System (SNDS, for *Système National des Données de Santé*) is a national-level inclusive data repository that was implemented in 2017 to facilitate the secondary use of the French administrative databases. The SNDS includes the national hospitalization discharge summary database, covering 65 million people with a longitudinal and exhaustive follow-up of patients regarding their hospital's history. The hospital discharge diagnoses have been recorded using ICD-10 since 2008[24]. We extracted all ICD-10 codes from all hospital discharge summaries between 2008 and 2017.

2.3 Definitions

We considered the 3-character ICD codes to reduce the number of codes while preserving a clinical or epidemiological interpretability. For example, the code C50.1 was standardized as C50. We extracted from the database tuples consisted of a timestamp (year-month), an ICD-10 code, the number of distinct patients annotated with the code for the given timestamp, the overall number of patients annotated with any code for the given timestamp. For each 3-character code and by the month, we computed the relative frequency, defined as follows:

$$Freq(t, ICD\ code) = \frac{\text{distinct number of patients with the ICD code } at\ time\ t}{\text{distinct number of patients with any code } at\ time\ t} \quad (1)$$

We applied a moving means, using the 6 months prior and after each point, to reduce the impact of extreme variations. To identify amplitude variations of the smoothed frequencies, we computed the relative amplitude defined as:

$$\text{Relative Amplitude(ICD code)} = \frac{|\min Freq(t,ICD\ code) - \max Freq(t,ICD\ code)|}{\text{Mean of min and max of } Freq(t,ICD\ code)} \quad (2)$$

2.4 Temporal Clustering

We included in the temporal clustering analysis only ICD codes with data covering the entire period of the study (2008-2017) to exclude the infrequent codes. We performed a temporal

hierarchical clustering analysis on the moving means time series. After normalizing the time series, we computed distances using the symbolic aggregate approximation (SAX) representation. Proposed by Lin *et al.*[26,27], the SAX representation is model-free and has been developed to deal with the heterogeneity of time series curves and pattern discovering. The SAX representation has two parameters: the amount of equal sized frames that the series will be reduced to (parameter w) and the size of the alphabet (parameter a, a>2). The SAX representation is performed in 2 steps: (1) the first step reduces the dimension of time series (2) the second step transforms the reduced time series into a symbolic representation. For the first step, each time series $C_{raw} = \left(c_{raw,1}, \dots, c_{raw,N}\right)$ of length N (N=120 months) is represented in a w-dimensional space by a vector $C_{SAX} = \left(c_{SAX,1}, \dots, c_{SAX,w}\right)$ defined as: for $i \in [\![1, w]\!]$,

$$c_{SAX,i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad (3)$$

The second step transforms the vector C_{SAX} into a vector $\hat{C}_{SAX} = \left(\widehat{c_{SAX,1}}, \dots, \widehat{c_{SAX,w}}\right)$ where $\forall j \in [\![1, w]\!], \widehat{c_{SAX,J}} \in alphabet$, $alphabet$ is the chosen alphabet ($Card(alphabet) = a$). The mapping between $c_{SAX,j}$ and $\widehat{c_{SAX,J}}$ is obtained as follows:

$$\widehat{c_{SAX,J}} = alphabet_j \; iif \; \beta_{j-1} \leq c_{SAX,j} < \beta_j \quad (4)$$

The set $\beta = \{\beta_j, j \in [\![1, a-1]\!]\}$ is obtained by partitioning the set of values in equiprobable regions under a Gaussian distribution assumption.

There is no recommendation for choosing the parameters (w, a). To adopt a fully unsupervised approach, we chose the combination of parameters (w, a) that maximizes the average silhouette[28]. We applied agglomerative hierarchical clustering on the distances matrix computed. We chose the number of clusters using the Silhouette method[29].

2.5 Statistical Software, Reproducibility and Ethical Considerations

All scripts are available at *https://github.com/equipe22/ICDtimeliness*. All analyses were performed with the R statistical software v.3.4.2. Temporal clustering was performed using the TSclust package[30]. This study was conducted under the methodological guidelines MR-005 of the French national data privacy authority[25].

3 RESULTS

3.1 Description of ICD codes

We extracted 238,334,751 encounters between 2008, January 1st and 2017, December 31st, corresponding to 56,621,773 distinct patients. The number of encounters per month slightly increased from 1,842,241 in January 2008 to 2,156,443 in December 2017. The number of patients per month slightly increased from 1,403,779 in January 2008 to 1,513,826 in December 2017. 1,006 ICD codes presented a variation of the relative amplitude of frequencies lower than 50%, 510 codes between 50% and 100% and 521 greater than 100%.

3.2 Temporal Clustering

Out of the 2,037 codes included in the study, we kept the 1,758 codes covering the entire period 2008-2017 for the clustering step. We extracted the moving means for the 1,758 ICD codes respecting the inclusion criteria. We identified the parameters of the SAX representation maximizing the silhouette: 20 frames (w), and alphabet size (a) of 6. The optimal number of clusters with the Silhouette and Elbow methods was 4. Figure 1 presents the dendrogram of the hierarchical clustering. Figure 2 presents the normalized moving mean time series for each cluster.

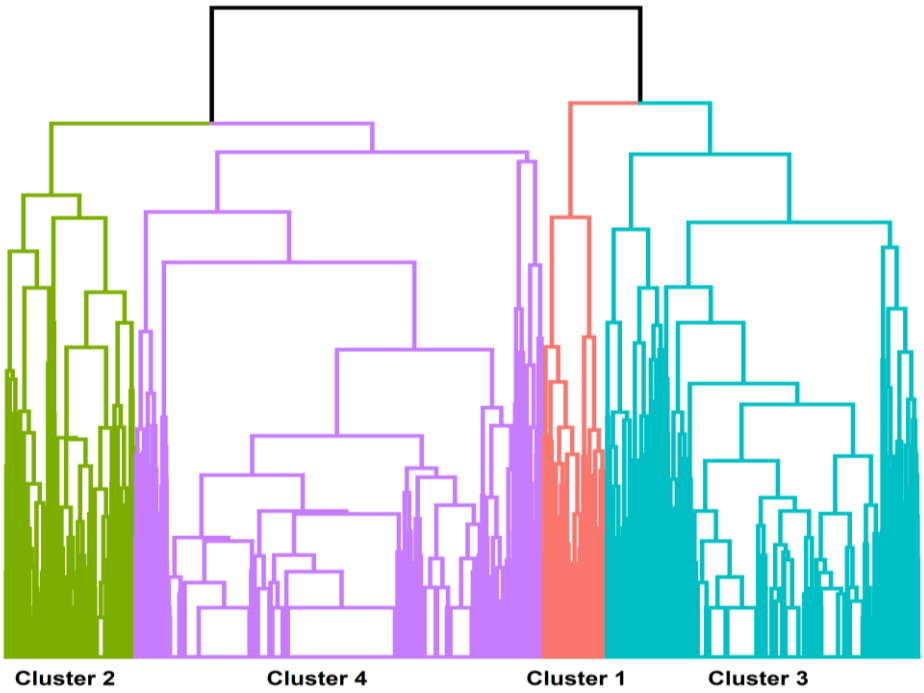


Figure 1: Dendrogram of the hierarchical clustering of the ICD normalized moving mean time series with the 4 selected clusters

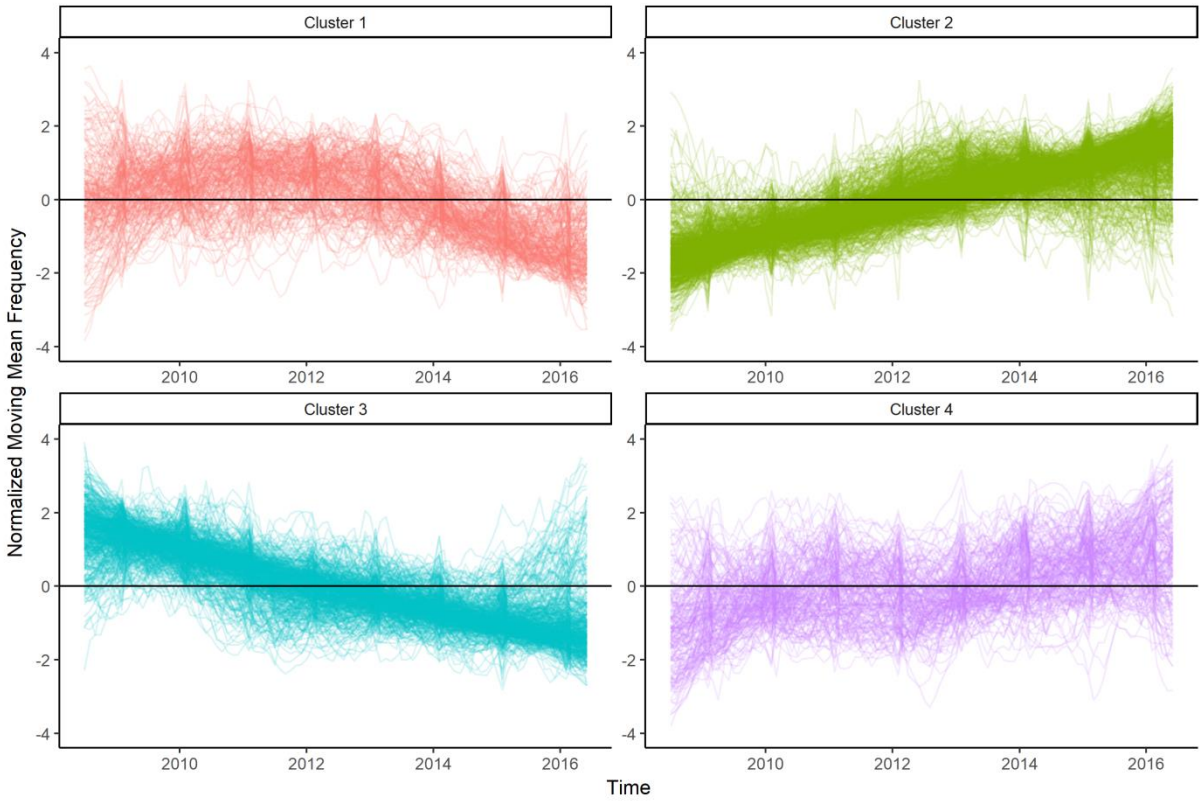
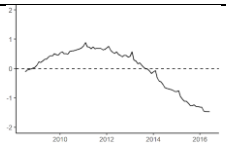
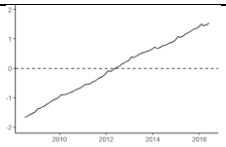
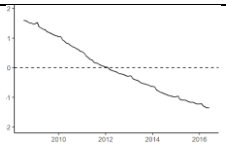
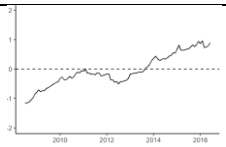


Figure 2: Representation of the ICD normalized moving mean time series by cluster

Table 1: Characteristics of clusters and examples of codes.

ICD-10 codes	**Cluster 1 (N=121)**	**Cluster 2 (N=249)**	**Cluster 3 (N=601)**	**Cluster 4 (N=787)**
Mean curve				
Average silhouette widths	0.250	-0.08	0.424	0.462
Examples *				
Infectious	B05	A53	A37	
Metabolic			E46	E43, E44
Cancer			C53, C55	C54

* A37: Whooping cough, A53: Syphilis, A54: Gonococcal infection, B05: Measles, C53: Malignant neoplasm of cervix uteri, C54: Malignant neoplasm of corpus uteri, C55: Malignant neoplasm of uterus, part unspecified, E43: Unspecified severe protein-energy malnutrition, E44: Protein-energy malnutrition of moderate and mild degree, E46: Unspecified protein-energy malnutrition.

3.3 Examples of evolution profiles

We selected ICD codes in each cluster to illustrate the evolution profiles. Figure 3 presents the evolution profiles of ICD codes for infectious diseases. Figure 4 presents the evolution profiles of metabolic disease and cancer codes.

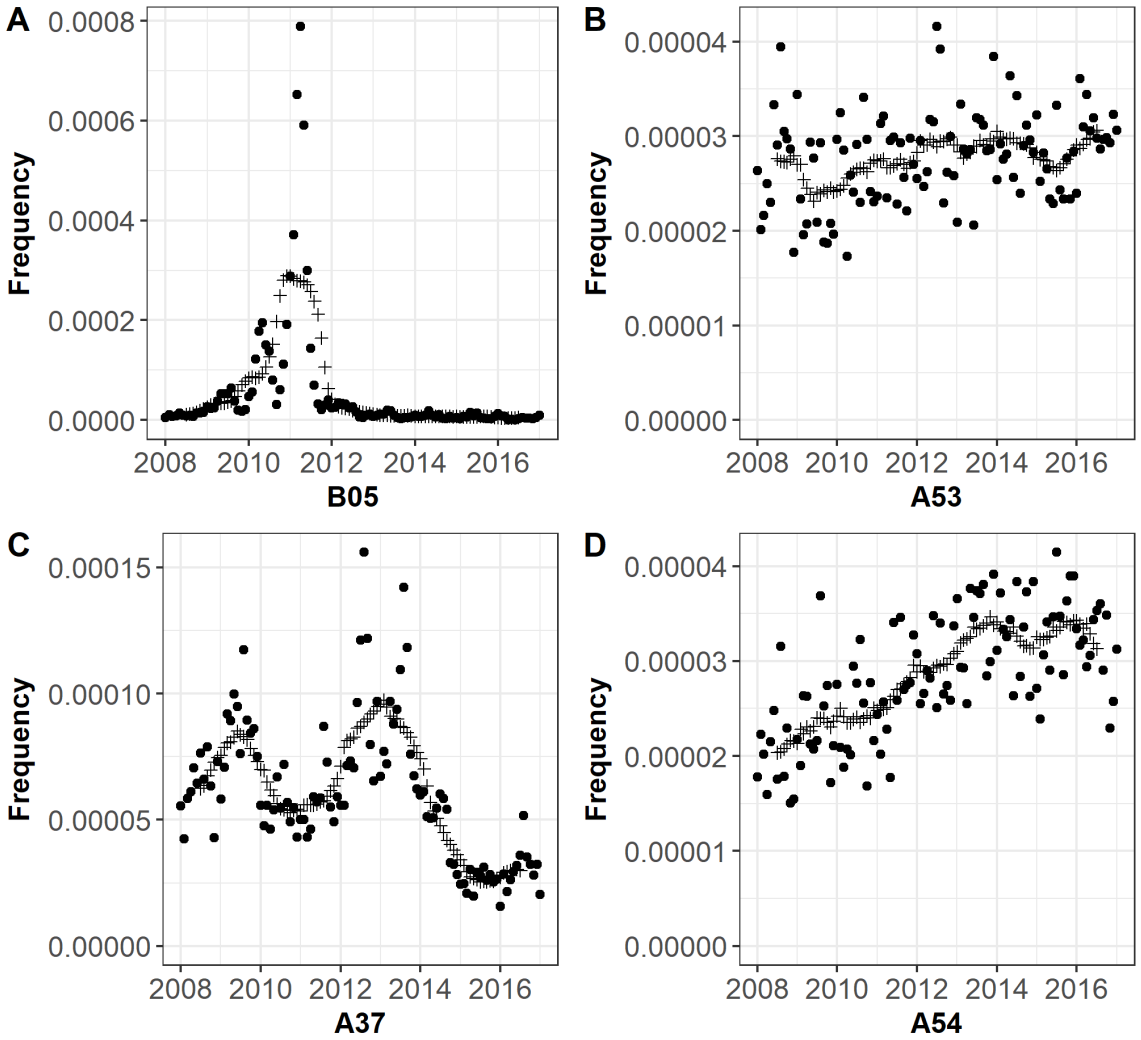


Figure 3: Evolution's profiles of examples of infectious codes. A37: Whooping cough, A53: Syphilis, A54: Gonococcal infection, B05: Measles. Circles represent the actual values, plus a smoothed value over 12 months.

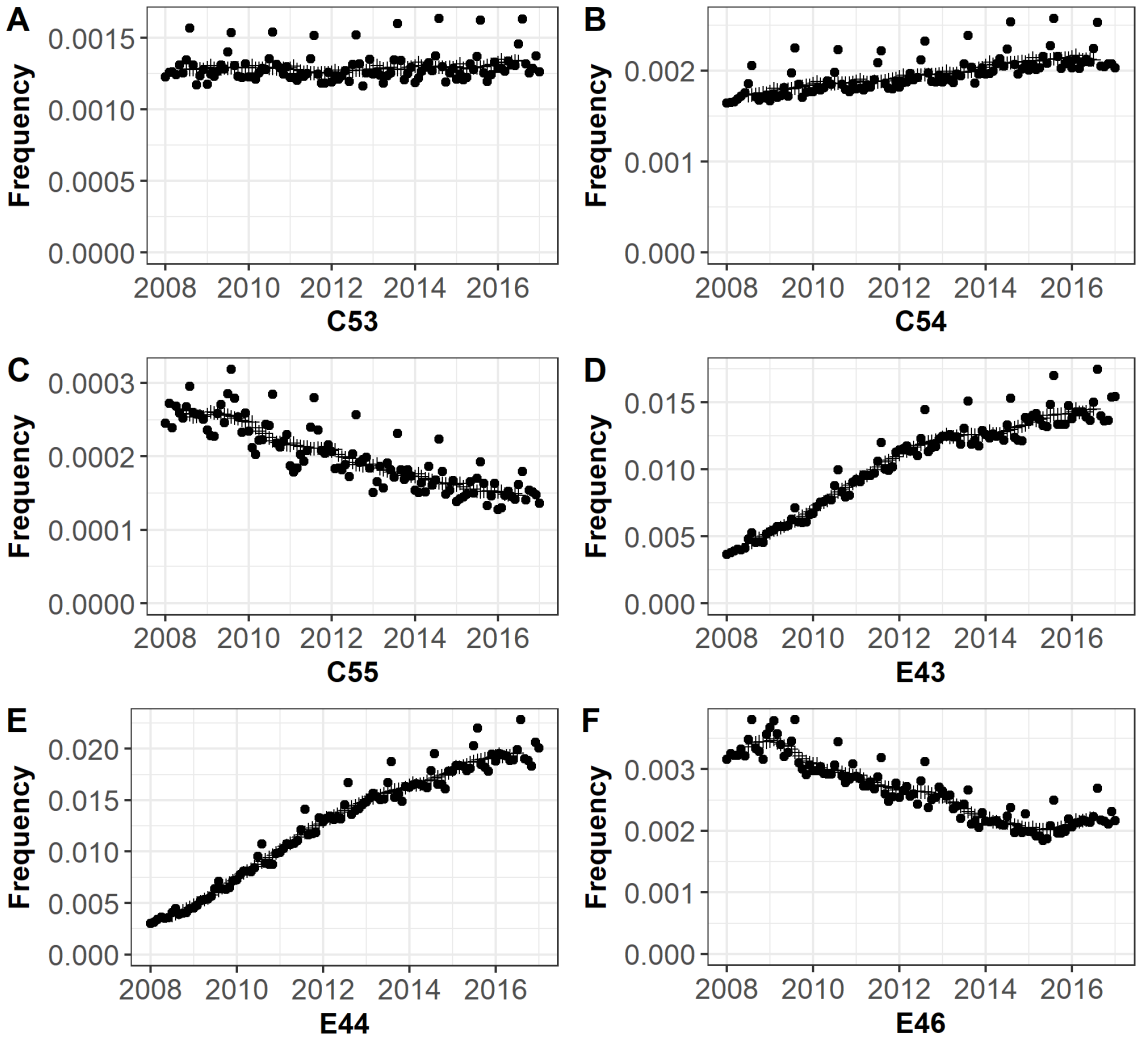


Figure 4 : Evolution profiles of examples of metabolic and cancer codes. C53: Malignant neoplasm of cervix uteri, C54: Malignant neoplasm of corpus uteri, C55: Malignant neoplasm of uterus, part unspecified, E43: Unspecified severe protein-energy malnutrition, E44: Protein-energy malnutrition of moderate and mild degree, E46: Unspecified protein-energy malnutrition. Circles represent the actual values, plus a smoothed value over 12 months.

4 DISCUSSION

4.1 Technical Findings

The overall results showed a strong instability (i.e. large variation of frequency over time) of the use of ICD codes over time, with an important variation of the relative amplitude of the frequencies: 510 codes have relative variations comprised between 50% and 100% and 521 greater than 100%. We observed 784 codes with increasing profiles (cluster 4) and 601 codes with decreasing profiles (cluster 3). We analyzed the evolution of three categories of ICD codes (infectious, metabolic and cancer). We described the evolution of ICD codes of infectious diseases to valid our methodology. Infectious diseases are monitored through epidemiological surveillance networks and the outbreaks are measured. In appendices, we detail how examples presented in figure 3 are related to epidemiological phenomena. Epidemiological outbreaks are detected with our 6-months smoothing. However, all trends cannot be clearly explained by epidemiological events.

External and intrinsic factors. We distinguished between external factors due to changes in billing, organization, policy or regulation and intrinsic factors due to epidemiological phenomenon. External factors can have a major influence and bias the results of longitudinal studies. In appendice, we explored two cases of such phenomena: the increase of malnutrition ICD codes (E43, E44 and E46) explained by the evolution of practices of coding optimization (for financial reason, an example of billing effect); and the evolution of codes of malignant neoplasms of the uterus explained by an improvement of precision of coding (learning effect). The billing effect is a change of coding practice due to financial incitation of the national reimbursement policy. The learning effect is an improvement of the precision of coding due to a better knowledge of the coding system by the physicians.

4.2 Significance for Data Reuse

International relevance. This study focuses on a national database in France. Other national databases could reflect similar external and intrinsic evolution. Integrated health information systems are being developed for both public health and research purposes[31]. Collaborative studies across data sources in several countries require explicit documentation of the external factors that may influence the distributions of the variables stored in the distributed databases. Our result suggests that a common terminology does not guarantee interoperability. Interoperability is affected by many factors including the potential link between reimbursement policy and coding system[32] or the change in coding system policy[33].

Analytic interoperability has to move beyond the traditional mapping approaches, where the emphasis is exclusively on aligning database schemas and codes and implement solutions in which the analysis of data can be ported between data sets, where although the codes may have been mapped the characteristics of the populations may lead to incompatible results if the same algorithms or statistical analyses are naively run 'as is' across datasets. As more research projects and surveillance methods will rely on reusing health administrative databases, these efforts will become crucial.

Impact on phenotyping. Most of the studies looking for associations between genotypes and phenotypes reuse ICD codes from EHRs as phenotype information[7]. The timeliness issue may have an impact on phenotyping algorithms and analyses distributed through networks of clinical data warehouses. Phenotype algorithm use structured and unstructured data to better identify cohorts of subjects within the health data. ICD codes are often an important source of structured data. For example, in PheKB[34], 70% of the phenotype algorithms use ICD codes. Therefore, our results suggest that developing phenotyping algorithms from administrative database alone, without taking the context of evolution of the code, can lead to substantial measurement bias in population studies.

4.3 Related Work

This study takes place in the context of temporal plausibility. The general topic of data quality has been widely explored in the literature[35,36]. Kahn *et al.*[24] have defined the plausibility as "features that describe the believability or truthfulness of data values." Kahn proposed a distinction between atemporal plausibility and temporal plausibility. The atemporal plausibility "seeks to determine if observed data values, distributions, or density agree with local or "common" knowledge (Verification) or from comparisons with external sources that are deemed to be trusted or relative gold standards (Validation)". Whereas the temporal plausibility "seeks to determine if time-varying variables change values as expected based on known temporal properties or across one or more external comparators or gold standards." Previous studies evaluated the quality of administrative databases in terms of validation studies for atemporal plausibility, e.g.,[37,38]. Quan *et al.*[37] studied the agreement between charts and ICD-9 codes on occurrence of procedures in patients admitted to either general medicine or general surgery services. They noticed a good specificity (greater than 99%) and an uncertain sensitivity (from 0 to 87.5%). They also analyzed positive predictive values, negative predictive values and kappa. This approach is classical in data quality analyses of administrative database. Hinds *et al.*[38] proposed a review of data quality studies of administrative health databases in Canada. Most of the data quality measures concern sensitivity (64.2%), specificity (55%), negative predictive value (43%), positive predictive value (58.3%) and kappa (30.5%). The coding errors were explained by various factors, such as clinicians' knowledge of terminologies, clinicians' experience in coding, diligence for compiling information, transcriber's ability to read notes and EHRs, and even intentional errors like underspecification and upcoding[39].

4.4 Remaining Challenges

Clusters and exploration of internal and external causes. In our study, we have adopted a fully unsupervised approach for temporal clustering to explore the existence of evolution profiles

without prior knowledge on the evolution. Other approaches have been developed, for example Sacchi *et al.*[40] proposed a two-step approach combining a qualitative representation of time series and a hierarchical clustering. A parametric approach based on piecewise regressions could provide another description for hypothesis-driven exploration. The main challenge resides in the identification of the origin of the evolution (intrinsic epidemiological or external factors). The detailed analysis of profiles show that the same cluster can contain profiles influenced by either cause and in some case, both causes simultaneously. In most cases, additional knowledge and sources are probably required to determine automatically the origin of the profile.

Contextualization. Saez *et al.*[22] have defined the contextualization as "the degree to which data is correctly/optimally annotated with the context in with it was acquired." We measured the stability of the frequencies and proposed a contextualization for a limited number of ICD-10 codes. The retrospective annotation of the timeliness of all the ICD codes is an important task and a perspective for a future work. We proposed some fundamental dimensions of the timeliness assessment. Other DQ metrics can be applied as the Information-Geometric Temporal or the Probability distribution function statistical process control[41]. Furthermore, we limited our analysis to the observation of the timeliness dimension. The DQ assessment of administrative database is broader than the timeliness dimension, for example, the national hospitalization database can be seen as multisources data and other DQ metrics like the source probabilistic outlyingness metric or the global probabilistic deviation metric could be used[42].

4.5 Recommendations

In order to address the different aspects of analytic harmonization, the health research community must develop a set of metadata. These metadata should represent this information in such a way as to facilitate comparison between data sources enabling the translation of research questions between these sources. While research efforts have developed shared

frameworks for provenance[43,44], our study demonstrates that we still need a framework for the capturing and representation of information about data quality. Data quality is a shared concern in research, and in public health[45]. Therefore, data quality profiling as metadata should be integrated in big data infrastructure as proposed by Merino *et al.*[46] in their "Data-Quality-in-Use model." Metadata definitions and ontologies for data sharing will enable the fingerprinting of data repositories and make the researchers aware of the quality of the available data. An IT framework for capturing detailed formal metadata about data sources, based on shared catalogs and automated data annotation would facilitate systematic annotation and harmonization.

5 CONCLUSION

We proposed an evaluation of the timeliness of ICD-10 codes in the French hospitalization summary discharges. External and intrinsic factors influencing the data quality of administrative data are heterogeneous and time-varying; human annotation remains necessary. A key challenge is to annotate these data repositories by producing machine-readable metadata to foster the improvement of the machine learning performances.

REFFERENCES

[1] N. Gavrielov-Yusim, M. Friger, Use of administrative medical databases in population-based research, J. Epidemiol. Community Health. 68 (2014) 283–287. doi:10.1136/jech-2013-202744.

[2] P. Tuppin, J. Rudant, P. Constantinou, C. Gastaldi-Ménager, A. Rachas, L. de Roquefeuil, G. Maura, H. Caillol, A. Tajahmady, J. Coste, C. Gissot, A. Weill, A. Fagot-Campagna, Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France, Rev. Epidemiol. Sante Publique. (2017). doi:10.1016/j.respe.2017.05.004.

[3] W. World Health Organization, International Classification of Diseases (ICD), (n.d.). http://www.who.int/classifications/icd/en/.

[4] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience, Int. J. Med. Inform. 102 (2017) 21–28. doi:10.1016/j.ijmedinf.2017.02.006.

[5] S.D. Grosse, S.L. Boulet, D.D. Amendah, S.O. Oyeku, Administrative Data Sets and Health Services Research on Hemoglobinopathies, Am. J. Prev. Med. 38 (2010) S557–S567. doi:10.1016/j.amepre.2009.12.015.

[6] A. Bello, R. Padwal, A. Lloyd, B. Hemmelgarn, S. Klarenbach, B. Manns, M. Tonelli, Using linked administrative data to study periprocedural mortality in obesity and chronic kidney disease (CKD), Nephrol. Dial. Transplant. 28 (2013) iv57-iv64. doi:10.1093/ndt/gft284.

[7] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D.R. Masys, D.M. Roden, D.C. Crawford, PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, Bioinformatics. 26 (2010) 1205–1210. doi:10.1093/bioinformatics/btq126.

[8] A. Neuraz, L. Chouchana, G. Malamut, C. Le Beller, D. Roche, P. Beaune, P. Degoulet, A. Burgun, M.-A. Loriot, P. Avillach, Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics, PLoS Comput. Biol. 9 (2013) e1003405. doi:10.1371/journal.pcbi.1003405.

[9] M. Yurkovich, J.A. Avina-Zubieta, J. Thomas, M. Gorenchtein, D. Lacaille, A systematic review identifies valid comorbidity indices derived from administrative health data, J. Clin. Epidemiol. 68 (2015) 3–14. doi:10.1016/j.jclinepi.2014.09.010.

[10] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L.D. Saunders, C.A. Beck, T.E. Feasby, W.A. Ghali, Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data, Med. Care. 43 (2005) 1130–1139. doi:10.1097/01.mlr.0000182534.19832.83.

[11] C. King, L. Garcia Alvarez, A. Holmes, L. Moore, T. Galletly, P. Aylin, Risk factors for healthcare-associated urinary tract infection and their applications in surveillance using hospital administrative data: a systematic review, J. Hosp. Infect. 82 (2012) 219–226. doi:10.1016/j.jhin.2012.05.004.

[12] R.M. Carnahan, R.A. Herman, K.G. Moores, A systematic review of validated methods for identifying transfusion-related sepsis using administrative and claims data, Pharmacoepidemiol. Drug Saf. 21 (2012) 222–229. doi:10.1002/pds.2322.

[13] J.A. Singh, J.A. Kundukulam, M. Bhandari, A systematic review of validated methods for identifying orthopedic implant removal and revision using administrative data, Pharmacoepidemiol. Drug Saf. 21 (2012) 265–273. doi:10.1002/pds.2309.

[14] C. Zhan, Administrative data based patient safety research: a critical review, Qual. Saf. Heal. Care. 12 (2003) 58ii–63. doi:10.1136/qhc.12.suppl_2.ii58.

[15] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers., Stud. Health Technol. Inform. 216 (2015) 574–8. http://www.ncbi.nlm.nih.gov/pubmed/26262116 (accessed September 7, 2017).

[16] OHDSI, (n.d.). http://ohdsi.org/.

[17] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.U. Prokosch, M. Sedlmayr, Towards Implementation of OMOP in a

German University Hospital Consortium, Appl. Clin. Inform. 09 (2018) 054–061. doi:10.1055/s-0037-1617452.

[18] S.T. Rosenbloom, R.J. Carroll, J.L. Warner, M.E. Matheny, J.C. Denny, Representing Knowledge Consistently Across Health Systems, Yearb. Med. Inform. 26 (2017) 139–147. doi:10.15265/IY-2017-018.

[19] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data. 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[20] D. Agniel, I.S. Kohane, G.M. Weber, Biases in electronic health record data due to processes within the healthcare system: retrospective observational study, BMJ. (2018) k1479. doi:10.1136/bmj.k1479.

[21] K. Mues, A. Liede, J. Liu, J. Wetmore, R. Zaha, B.D. Bradbury, A. Collins, D. Gilbertson, Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US, Clin. Epidemiol. Volume 9 (2017) 267–277. doi:10.2147/CLEP.S105613.

[22] C. Sáez, J. Martínez-Miranda, M. Robles, J.M. García-Gómez, Organizing data quality assessment of shifting biomedical data., Stud. Health Technol. Inform. 180 (2012) 721–5. http://www.ncbi.nlm.nih.gov/pubmed/22874286 (accessed April 3, 2018).

[23] G. Rey, A. Aouba, G. Pavillon, R. Hoffmann, I. Plug, R. Westerling, E. Jougla, J. Mackenbach, Cause-specific mortality time series analysis: a general method to detect and correct for abrupt data production changes, Popul. Health Metr. 9 (2011) 52. doi:10.1186/1478-7954-9-52.

[24] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, L. Schilling, A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes). 4 (2016) 18. doi:10.13063/2327-9214.1244.

[25] CNIL, Délibération n° 2018-256 du 7 juin 2018 portant homologation d'une méthodologie de référence relative aux traitements de données nécessitant l'accès par des établissements de santé et des fédérations aux données du PMSI, 2018.

[26] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proc. 8th ACM SIGMOD Work. Res. Issues Data Min. Knowl. Discov. - DMKD '03, ACM Press, New York, New York, USA, 2003: p. 2. doi:10.1145/882082.882086.

[27] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Discov. 15 (2007) 107–144. doi:10.1007/s10618-007-0064-z.

[28] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[29] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[30] P. Montero, J.A. Vilar, TSclust : An R Package for Time Series Clustering, J. Stat. Softw. 62 (2014) 1–43. doi:10.18637/jss.v062.i01.

[31] A. Burgun, E. Bernal-Delgado, W. Kuchinke, T. van Staa, J. Cunningham, E. Lettieri, C. Mazzali, D. Oksen, F. Estupiñan, A. Barone, G. Chène, Health Data for Public Health: Towards New Ways of Combining Data Sources to Support Research Efforts in Europe, Yearb. Med. Inform. 26 (2017) 235–240. doi:10.15265/IY-2017-034.

[32] L. Manchikanti, A.D. Kaye, V. Singh, M. V Boswell, The Tragedy of the Implementation of ICD-10-CM as ICD-10: Is the Cart Before the Horse or Is There a Tragic Paradox of Misinformation and Ignorance?, Pain Physician. 18 (n.d.) E485-95. http://www.ncbi.nlm.nih.gov/pubmed/26218946.

[33] C.A. Panozzo, T.S. Woodworth, E.C. Welch, T.-Y. Huang, Q.L. Her, K. Haynes, C. Rogers, T.J. Menzin, M. Ehrmann, K.E. Freitas, N.R. Haug, S. Toh, Early impact of the ICD-10-CM transition on selected health outcomes in 13 electronic health care databases in the United States, Pharmacoepidemiol. Drug Saf. (2018). doi:10.1002/pds.4563.

[34] J.C. Kirby, P. Speltz, L. V Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, S.B. Ellis, T. Lingren, W.K. Thompson, G. Savova, J. Haines, D.M. Roden, P.A. Harris, J.C. Denny, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Informatics Assoc. 23 (2016) 1046–1052. doi:10.1093/jamia/ocv202.

[35] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Informatics Assoc. 20 (2013) 144–151. doi:10.1136/amiajnl-2011-000681.

[36] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of EHR: Data quality issues and informatics opportunities, in: AMIA Summits Transl. Sci. Proc., 2010: pp. 1–5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041534/pdf/amia-s2010_cri_001.pdf (accessed August 23, 2017).

[37] H. Quan, G.A. Parsons, W.A. Ghali, Validity of Procedure Codes in International Classification of Diseases, 9th revision, Clinical Modification Administrative Data, Med. Care. 42 (2004) 801–809. doi:10.1097/01.mlr.0000132391.59713.0d.

[38] A. Hinds, L.M. Lix, M. Smith, H. Quan, C. Sanmartin, Quality of administrative health databases in Canada: A scoping review, Can J Public Heal. 107 (2016) 56. doi:10.17269/cjph.107.5244.

[39] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, Measuring Diagnoses: ICD Code Accuracy, Health Serv. Res. 40 (2005) 1620–1639. doi:10.1111/j.1475-6773.2005.00444.x.

[40] L. Sacchi, R. Bellazzi, C. Larizza, P. Magni, T. Curk, U. Petrovic, B. Zupan, TA-clustering: Cluster analysis of gene expression profiles through Temporal Abstractions, Int. J. Med. Inform. 74 (2005) 505–517. doi:10.1016/j.ijmedinf.2005.03.014.

[41] C. Sáez, O. Zurriaga, J. Pérez-Panadés, I. Melchor, M. Robles, J.M. García-Gómez, Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories, J. Am. Med. Informatics Assoc. 23 (2016) 1085–1095. doi:10.1093/jamia/ocw010.

[42] C. Sáez, M. Robles, J.M. García-Gómez, Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances, Stat. Methods Med. Res. 26 (2017) 312–336. doi:10.1177/0962280214545122.

[43] V. Curcin, E. Fairweather, R. Danger, D. Corrigan, Templates as a method for implementing data provenance in decision support systems, J. Biomed. Inform. 65 (2017) 1–21. doi:10.1016/j.jbi.2016.10.022.

[44] S. Xu, T. Rogers, E. Fairweather, A. Glenn, J. Curran, V. Curcin, Application of Data Provenance in Healthcare Analytics Software: Information Visualisation of User Activities., AMIA Jt. Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci. 2017 (2018) 263–272. http://www.ncbi.nlm.nih.gov/pubmed/29888084.

[45] H. Chen, D. Hailey, N. Wang, P. Yu, A Review of Data Quality Assessment Methods for Public Health Information Systems, Int. J. Environ. Res. Public Health. 11 (2014) 5170–5207. doi:10.3390/ijerph110505170.

[46] J. Merino, I. Caballero, B. Rivas, M. Serrano, M. Piattini, A Data Quality in Use model for Big Data, Futur. Gener. Comput. Syst. 63 (2016) 123–130. doi:10.1016/j.future.2015.11.024.

APPENDICE: Looking for Extrinsic and Intrinsic Factors Explaining the Observed Evolution of the Examples of Figures 3 and 4

Figure 3A (B05, Measles). The prevalence of measles is close to zero and stable over time, except around 2011 where an explosive peak is detected. More precisely, we observed three peaks of increasing intensity between 2009 and 2012. As measles is a notifiable disease in France, cases are exhaustively reported in national registers. Based on these national surveillance data, Antona *et al.*[1,2] reported that France experienced a massive measles outbreak in 2010-2011, accounting for more than half of the 30,000 cases in Europe during this period, with almost 5,000 persons hospitalized. The epidemic curve of their article showed that the number of cases started increasing in mid-2008, evolving in 3 epidemic waves: 2009, 2010 and the highest epidemic waves in 2011. Our results are consistent with this data.

*[1] Antona D, Lévy-Bruhl D, Baudon C, Freymuth F, Lamy M, Maine C. Measles elimination efforts and 2008-2011 outbreak, France. Emerg Infect Dis. 2013;**19**(3):357-64.* DOI: 10.3201/eid1903.121360 PMID: 23618523

[2] Institut de Veille Sanitaire (InVS). Épidémie de rougeole en France. Actualisation des données de surveillance au 13 mars 2017. [Measles outbreak in France. Updated surveillance data on 13 Mar 2017]. Saint Maurice: InVS.. French.

Available from: http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-infectieuses/Maladies-a-prevention-vaccinale/Rougeole/Points-d-actualites/Archives/Epidemie-de-rougeole-en-France.-Actualisation-des-donnees-de-surveillance-au-13-mars-2017

Figure 3B (A53, Syphilis). The prevalence of syphilis slightly increased since 2010. In France, the *RésIST* network contributes to monitoring sexually transmitted infections such as Syphilis

or Gonorrhea. The collection of demographic, clinical, biological and behavioral data is based on different diagnostics, information and screening sites, and medical consultations.

Surveillance data from RésIST [3] confirmed our observed trend, which translates the increase in risky sexual behaviors, in particular in men who have sex with men. Despite the heterogeneity of surveillance systems, European surveillance data lead to the same findings [4].

[3] Ndeindo Ndeikoundam Ngangro, Delphine Viriot, Nelly Fournet, Bertille De Barbeyrac, Agathe Goubard, Nicolas Dupin, Michel Janier, Isabelle Alcaraz, Michel Ohayon, Nathalie Spenatto, Chantal Vernay-Vaisse, les référents des Cire, Josiane Pillonel, Florence Lot. Bacterial sexually transmitted infections in France : recent trends and characteristics in 2015. Bull Epidemiol Hebd. N° 41-42.

Available from: http://opac.invs.sante.fr/index.php?lvl=notice_display&id=13182

[4] European Centre for Disease Prevention and Control. Sexually transmitted infections in Europe 2013. Stockholm: ECDC; 2015. 124 p.

Available at: http://ecdc.europa.eu/en/publications/Publications/sexual-transmitted-infections-europe-surveillance-report-2013.pdf

Figure 3C (A37, Whooping cough). On Figure 3C, we observe two outbreaks in 2009 and 2012. After 2012, the prevalence of whooping cough decreased until 2016. In France, *RENACOQ*, a sentinel hospital-based voluntary surveillance network has been established in 1996, and covers about 30% of hospitalized pertussis pediatric cases. Surveillance data from RENACOQ [5,6] confirmed our observed outbreaks: there have been increases in pertussis prevalence in the past few years in France, and in particular, two epidemic peaks occurred in 2009 and 2012.

[5] Tubiana S, Belchior E, Guillot S, Guiso N, Lévy-Bruhl D; Renacoq Participants. Monitoring the Impact of Vaccination on Pertussis in Infants Using an Active Hospital-based

Pediatric Surveillance Network: Results from 17 Years' Experience, 1996-2012, France. Pediatr Infect Dis J. 2015 Aug;34(8):814-20. doi: 10.1097/INF.0000000000000739.

[6] Principales caractéristiques des cas de coqueluche identifiés par le réseau Renacoq, 1996-2015. Réseau RenaCoq. InVS.

Available at: http://invs.santepubliquefrance.fr/fr/content/download/31035/158002/version/7/file/Tableau_cas_coqueluche_+1996-2015.pdf

Figure 3D (A54, Gonococcal infection). Gonococcal infections highly increased between 2008 to 2017. Surveillance data from RésIST, and Renago, a laboratory network that collects demographic and biological data for gonorrhea, show similar results. The diffusion of multi-resistant strains in a context of more and more frequent transmission of gonococci explains these trends[7,8].

[7] La Ruche G, Goubard A, Bercot B, Cambau E, Semaille C, Sednaoui P. Gonococcal infections and emergence of gonococcal decreased susceptibility to cephalosporins in France, 2001 to 2012. Euro Surveill. 2014;19(34). pii: 20885. Available at: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20885

[8] Mohammed H, Ison CA, Obi C, Chisholm S, Cole M, Quaye N, et al; GRASP Collaborative Group. Frequency and correlates of culture-positive infection with Neisseria gonorrhoeae in England: a review of sentinel surveillance data. Sex Transm Infect. 2015;91(4):287-93.

Figures 4A, 4B and 4C (C53, Malignant neoplasm of cervix uteri; C54, Malignant neoplasm of corpus uteri; C55, Malignant neoplasm of uterus, part unspecified). The C53 code (Malignant neoplasm of cervix uteri) showed a stationary trend whereas French and European literature data reported a decreasing incidence over the past years, mainly due to the widespread

implementation of screening programs (cervical smear tests) since the 1960s and HPV vaccination since 2006 in France[9-11]. Conversely, the C54 code (Malignant neoplasm of corpus uteri) increase slightly during the study period, in accordance with epidemiological data [9,10,12]. The unspecified code C55 (Malignant neoplasm of uterus, part unspecified) shows a 2-fold decrease between 2008 and 2017. By analyzing the curves together, we can suppose that the stationary trend of C53 may be explained by an improvement of coding precision, leading to abandon the C55 code. This billing effect masked the real cervix cancer evolution over time. To analyze time trends of uterine cancer incidence with data extracted from the ICD-10 classification, we had to deal with varying coding precision over time. Loos *et al.* focused on imprecisely coded uterine cancer deaths based on ICD-10 codes and developed a reallocation procedure for the "unspecified" category. This method may be added as a supplementary step in incidence data treatment based on ICD codes to avoid misleading interpretation[13].

[9] Jéhannin-Ligier K, Dantony E, Bossard N, Molinié F, Defossez G, Daubisse-Marliac L, Delafosse P, Remontet L, Uhry Z. Projection de l'incidence et de la mortalité par cancer en France métropolitaine en 2017. Rapport technique. Saint-Maurice : Santé publique France, 2017. 80 p.

Available from: http://invs.santepubliquefrance.fr/Publications-et-outils/Rapports-et-syntheses/Maladies-chroniques-et-traumatismes/2018/Projection-de-l-incidence-et-de-la-mortalite-par-cancer-en-France-metropolitaine-en-2017

[10] Binder-Foucard F, Bossard N, Delafosse P, Belot A, Woronoff A.-S, Remontet L the French network of cancer registries (Francim). Cancer incidence and mortality in France over the 1980–2012 period: Solid tumors. Revue d'épidémiologie et de Santé Publique 62 (2014) 95–108. http://dx.doi.org/10.1016/j.respe.2013.11.073

[11] Knudsen A, Schledermann D, Nyvang GB, Mogensen O, Herrstedt J. Trends in gynecologic cancer among elderly women in Denmark, 1980-2012.Acta Oncol. 2016;55 Suppl 1:65-73. doi: 10.3109/0284186X.2015.1115119. Epub 2016 Jan 19.

[12] Lortet-Tieulent J, Ferlay J, Bray F, Jemal A. International Patterns and Trends in Endometrial Cancer Incidence, 1978–2013. JNCI J Natl Cancer Inst (2018) 110(4): djx214. doi: 10.1093/jnci/djx214

[13] Loos AH, Bray F, McCarron P, et al. Sheep and goats: separating cervix and corpus uteri from imprecisely coded uterine cancer deaths, for studies of geographical and temporal variations in mortality. Eur J Cancer 2004;40: 2794-803.

Figures 4D, 4E and 4F (E43, Unspecified severe protein-energy malnutrition; E44, Protein-energy malnutrition of moderate and mild degree; E46, Unspecified protein-energy malnutrition). In figure 4D, 4E, 4F we explored the E43, E44, E46 malnutrition codes. These codes are often used in France to demonstrate the severity of a hospital stay and their coding is therefore used to optimize reimbursement by hospitals. The E43 code (Unspecified severe protein-energy malnutrition) and the E44 code (Protein-energy malnutrition of moderate and mild degree) show a global increase. Considering these results, malnutrition during hospital stays would have increased by a factor of 6 between 2008 and 2017. There are no available arguments that can justify these trends. Moreover, undercoding of malnutrition has been highlighted in France as early as 2003[14]. In recent years, practices of coding optimization have been introduced in many institutions[15,16]. Therefore, coding optimization may explain the decreasing trend of E46 (Unspecified protein-energy malnutrition) for the benefit to E43 and E44 codes that entailed a gain of precision coding. The figure 4D (E43) presents a reduced slope starting in 2012, probably because financial optimization was reached globally, and no further improvement could be achieved.

[14] Zazzo J-F. Programme national nutrition santé. Dénutrition : une pathologie méconnue en société d'abondance. Ministère de la santé. http://www.sante.gouv.fr/IMG/pdf/brochure_denutrition.pdf

[15] Strukov A, Perozziello A, Delon M, Picouleau D, Buzzi JC. The improvement of coding with paramedical information. Journal de Gestion et d'Économie Médicales 2016, Vol. 34, n° 5-6, 263-273

[16] Estran S, Tifratene K, Barthélémi C, PM Roger, Hebuterne X, Schneider S. Déterminants du défaut de codage de la dénutrition dans un service de médecine. Nutrition Clinique et Métabolisme Volume 30, Issue 2, June 2016, Page 123.