Information quality measurement of medical encoding support based on usability

John Puentes, Julien Montagner, Laurent Lecornu, Jean Michel Cauvin

Information quality measurement of medical encoding support based on usability

John Puentes[a,*], *Julien Montagner*[a], *Laurent Lecornu*[a], *Jean-Michel Cauvin*[b,c]

[a] *Institut Mines-Telecom; Telecom Bretagne, UEB; Dpt Image et Traitement de l'Information, Brest, France*
[b] *Département d'Information Médicale CHRU, Brest, France*
[c] *Inserm UMR 1101 LaTIM, Brest, France*

Keywords:
Information quality assessment
Medical information analysis
Clinical encoding
Usability
Decision support
Hospital information systems

Medical encoding support systems for diagnoses and medical procedures are an emerging technology that begins to play a key role in billing, reimbursement, and health policies decisions. A significant problem to exploit these systems is how to measure the appropriateness of any automatically generated list of codes, in terms of fitness for use, i.e. their quality. Until now, only information retrieval performance measurements have been applied to estimate the accuracy of codes lists as quality indicator. Such measurements do not give the value of codes lists for practical medical encoding, and cannot be used to globally compare the quality of multiple codes lists. This paper defines and validates a new encoding information quality measure that addresses the problem of measuring medical codes lists quality. It is based on a usability study of how expert coders and physicians apply computer-assisted medical encoding. The proposed measure, named ADN, evaluates codes Accuracy, Dispersion and Noise, and is adapted to the variable length and content of generated codes lists, coping with limitations of previous measures. According to the ADN measure, the information quality of a codes list is fully represented by a single point, within a suitably constrained feature space. Using one scheme, our approach is reliable to measure and compare the information quality of hundreds of codes lists, showing their practical value for medical encoding. Its pertinence is demonstrated by simulation and application to real data corresponding to 502 inpatient stays in four clinic departments. Results are compared to the consensus of three expert coders who also coded this anonymized database of discharge summaries, and to five information retrieval measures. Information quality assessment applying the ADN measure showed the degree of encoding-support system variability from one clinic department to another, providing global evaluation of quality measurement trends.

1. Introduction

The fundamental goal of medical encoding is to identify diagnosis related groups of patients and determine the corresponding healthcare expenses, billing, and reimbursement. Medical encoding is used in addition to record diseases morbidity and causes of mortality. This encoded information has become increasingly important, given its impact on medical activities evaluation at various levels of health

* *Corresponding author*. Tel.: +33 2290013 39.
E-mail address: John.Puentes@telecom-bretagne.eu (J. Puentes).

organizations. Moreover, encoding relevance affects patient management, along with epidemiologic, safety, research, and health policies decisions [1]. Medical codes are assigned to define diagnoses and procedures of each care episode that occurred during an inpatient stay. Codes represent [2,3]: main and secondary diagnoses, complications, comorbidities, primary and secondary procedures. Currently, most of medical encoding is carried out in two distinct manual manners:

- By expert coders who, without having any particular ancillary knowledge of the specific patient history, produce lists of codes that are considered to be exhaustive.
- By physicians who code essential aspects of the care episodes having some knowledge of the specific patient history, but usually generate a subset of the codes list produced by expert coders, because of practical restrictions (mainly focus on current diseases, limited awareness of encoding guidelines, and short available time).

In both cases medical encoding is expensive, taking much more time for the experts than for the physicians. The main reason is that, in addition to the patient record human coders have to examine hundreds of candidate codes in encoding references, to define the codes list that represents a given inpatient stay. Nevertheless, the pertinence of resulting codes sets strongly depends on the variable coders' expertise, which often produces inacceptable results like under or over encoding [1,4].

An emerging alternative is computer-assisted medical encoding technology. It analyses available patient information to automatically generate a list of most pertinent medical codes. Thereafter, coders select the appropriate codes corresponding to a specific inpatient stay. In more than a decade, several approaches have been developed to produce the corresponding encoding support lists, using varied types of input information. These studies considered, among others: extraction of semantic labels from documents [5,6]; matching of structured encoding forms and parsed clinical information [7]; correlation with precedent encoding results [8]; use of encoding rules [9]; codes linked to specific keywords [10]; a combination of an encoding classification with ontologies and natural language [11].

A significant problem with this kind of technology is how to measure the proposed codes lists appropriateness in terms of fitness for use, i.e. quality, according to: the distribution of correct and incorrect codes along the list, the amount of expected correct codes, the observation windows, and the variable list length. That information quality measurement should assert the practical value of any codes list, in a suitable manner adapted to the different encoding practices of hospitals and countries [12–15]. Automatically generated codes lists represent nevertheless information of variable quality, depending on the quality of input data. Such quality is complex to determinate particularly on the heterogeneous data sets of any hospital information system (HIS). Otherwise, the pertinence of codes lists produced by an encoding support system is conventionally estimated by comparing suggested codes, with a reference encoding done by an expert coder. Nevertheless, this approach does not provide any clues about the lists' value for medical encoding. Information retrieval performance measurements have been also used to estimate accuracy as quality indicator. These measurements do not give either the value of codes lists in the sense of their adequacy to encoding practices, and cannot be used to globally compare the quality of multiple codes lists. This paper thus addresses the problem of how to measure the appropriateness of an automatically generated codes list, in terms of fitness for use. We define and validate a new information quality measure that copes with limitations of previously applied measures. It is based on how expert coders and physicians make use of computer-assisted medical encoding. The proposed measure, named ADN, evaluates codes Accuracy, Dispersion, and Noise in the whole generated list, independently of its content and length. According to the ADN measure, the information quality of a codes list is fully represented by a single point within a normalized triangular space, partitioned by iso-quality lines. Moreover, our approach is reliable to examine and compare, using a unique scheme, the information quality of hundreds of codes lists, showing their practical value for encoding.

1.1. *Background*

In general, data quality analyses have been focused on well-known issues (wrong, missing or unusable data) produced by both humans and systems, at any stage of the data existence cycle. These issues accumulate generating varied complex functional problems [16,17]. Even if multiple approaches have been proposed to systematically identify, characterize, and correct inconsistencies produced by deficient data [18–22], information quality assessment remains a central and unexplored challenge [23–26]. Furthermore, encoding information quality measurement is particularly necessary when a regular audit of the associated HIS applications cannot be done, due to functional constraints.

Examined documents to produce medical codes vary from the whole patient record to discharge summaries. Included data and information are expected to be accurate, i.e. truly represent the element each value was intended for. As a consequence, only accuracy has been commonly considered as analogous to quality in the medical domain. For more than 30 years precision (*Pr*) and recall (*Rc*) have been the main applied information retrieval measures [27,28], along with complementary related evaluations [29,30]. For any retrieval system, its overall performance is comparatively determined using a set of precision-recall curves [31]. Rank measures of relevant documents can be calculated considering average weighed precision, by means of R-Measure and Q-measure [32,33]. Additionally, to handle incomplete information, binary relevance judgment defines globally a preference relation with respect to relevant documents [34]. Until now, only measurements used for information retrieval performance evaluation have been applied to estimate medical data and information quality. Based on some of these approaches, accuracy of computer-based patient records data was estimated applying two complementary measures [35]: *Cr* – correctness (proportion of correct data) and *Cm* – completeness (proportion of rightly recorded data), defined as:

$$Cr = \frac{tp}{tp + fp} = Pr = PPV \tag{1}$$

$$Cm = \frac{tp}{tp + fn} = Rc = Ss \tag{2}$$

where tp represents the correct recorded data (true positives), fp the wrongly recorded data (false positives) and fn the missing information (false negatives). These two measurements are equivalent, respectively, to Pr and Rc, or the positive predicted value (PPV) and sensitivity (Ss) criteria applied in information retrieval performance evaluation. Alternatively, the so-called F-measure (Fm) calculates the harmonic mean of precision and recall:

$$Fm = \frac{2(Pr \cdot Rc)}{Pr + Rc} \tag{3}$$

Otherwise, specific information quality measures for medical encoding lists have not been proposed in the literature. For that reason existing studies aiming at measuring the information quality of medical codes basically rely on the information retrieval concept of data accuracy. For instance, precision–recall was applied with limited results, to estimate the pertinence of encoding support compared to human coders [23]. Encoding quality of obstetric discharge data was evaluated applying correctness and completeness, to compare coded hospital complications with recoding of the same elements done by experts [36]. Individual and combined correctness and completeness indicators were used to assess human coders output with and without encoding support, applying the F-measure [37]. Manual encoding accuracy of causes-of-death was estimated taking into account, besides accuracy, the amount of required codes that were not originally included [38]. Morbidity encoding quality in general practice was assessed combining correctness and completeness evaluation to validate the construction of computerized medical records [39]. Other attempts have been made to estimate medical encoding information quality, intended to extend the accuracy notion, by: integrating prevalence, sensitivity, positive predictive value, and kappa statistic [40]; measuring the accuracy of automatically extracted diagnoses codes using a semantic distance [41]; determining the credibility of precision-recall [42]; and using the percentage of encoding errors combined to the assessment of codes sensitivity-specificity to detect encoding problems [14,43]. However, accuracy does not measure codes lists suitability for practical encoding. Moreover, none of these studies dealt with information quality of encoding support lists in terms of usability appropriateness.

1.2. *Limitations to measure medical encoding quality*

Information quality measurements based on precision-recall encompass two considerable limitations, when applied to similar lists of computer-assisted medical encoding systems:

- Unless obtained precision-recall curves are clearly separated by diverging trends and have the same length, it is not possible to assert which one has better encoding quality. For instance, when curves intersect each other in cases 1–2 and 3–4 of Fig. 1.
- Similarly, when lists have the same amount of proper codes but ordered differently, it is unclear which one fits better human coders' activity, like in cases 3–4 and 5–6 of Fig. 1.

As a result, these limitations impede to determine quantitatively, or visually on the 2D curves of the diagram, which of the codes lists in each pair is more appropriate under practical encoding conditions. Hence, usability pertinence remains unachievable from that perspective. Furthermore, this problem becomes extremely complex when hundreds of codes lists are compared to evaluate an encoding support system, confirming the impossibility to determine the usability pertinence of each list.

This paper has been divided in four other sections. Section 2 describes the encoding support system and data, the elements of the usability study on which is based this work, and defines the proposed ADN encoding information quality model. Section 3 first demonstrates the theoretical features of the ADN quality measure using simulated codes lists. It then looks at how the dominant approaches of the literature compare to the ADN measure, and presents the validation using simulated and real data. Results are discussed in Section 4. Section 5 outlines the contribution and perspectives of our work.

2. Materials and methods

Medical codes lists used in this work were automatically generated by an encoding support system developed and deployed at the university hospital CHRU Brest, France (Section 2.1). Since encoding information quality measurement greatly depends on how coders make use of encoding support lists, a usability study was first conducted (Section 2.2). A group of 6 coders (3 specialized expert coders and 3 experienced physicians) was requested to use during 5 months the previously indicated encoding support system, and then explain sequentially how they made use of it. The most significant findings of the system usability study serve to define the ADN quality measure model (Section 2.3).

2.1. *Encoding support system and data*

To generate medical codes lists the encoding support system processes the outputs of six HIS applications. Depending on these information sources, three information processing tools (Fig. 2) are used to generate either: semantic labels, estimation of pertinent codes, or probabilities. These results are aggregated considering their relative significance, to generate a unique list of ranked codes for a specific inpatient stay. This encoding support list is then examined by the physician to select appropriate codes. The three information processing tools which process documents stored in the HIS to produce medical codes are based on: the link between laboratory results and potential diagnoses by means of linguistic labels [44]; the evaluation of previously assigned codes to identify frequency patterns [45]; and the study of probabilistic relations between diagnoses codes and procedural parts of discharge summaries [46]. Results fusion integrates the partial heterogeneous information extracted

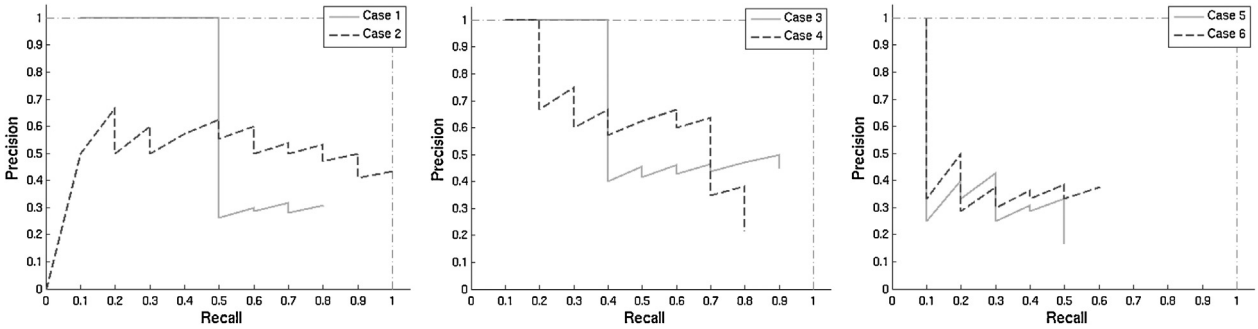

Fig. 1 – Compared evolution examples of precision-recall for different pairs of comparable codes lists, depending on their length.

by each method, generating one list of codes per inpatient stay [47]. Because of the heterogeneous nature and quality of exploited information, the encoding support output is affected in various manners: inappropriate code ranking, missing codes, dispersion of correct codes among incorrect ones, and noise induced by long sequences of incorrect codes. Certainly, expert coders and physicians may be able to rule out autonomously most of input information heterogeneity, applying learned heuristics and/or having access to additional patient documents. Encoding-support systems cannot solve those problems in an equivalent manner, making necessary to measure the information quality of each generated codes list.

Codes lists were generated using information collected from the HIS during 1 year. Information processed by the encoding system consisted of laboratory results, discharge summaries, administrative documents, and encoding reference (International Statistical Classification of Diseases and Related Health Problems, compiled by the World Health Organization [48]). Generated codes lists correspond to 502 anonymized discharge summaries of different inpatient stays, at four clinical departments (traumatology – 198 lists, obstetrics – 141 lists, urology – 105 lists, and cardiology – 58 lists). Generated coding information represents varying: length of stay (2–144 days), quantity of diagnoses per inpatient stay (1–21), and types of diagnosis codes (a total of 416 codes). Ground truth to verify the pertinence of encoding support quality was provided by consensus of the three expert coders who coded manually the same information processed by the encoding system. Whereas the encoding system generated automatically the corresponding lists of plausible codes in few seconds, several weeks of analysis were necessary for the expert coders. Otherwise, simulated sets of codes lists were used to evaluate the representation space boundaries, as well as the influence of variable size dispersion and noise windows of the quality measure model.

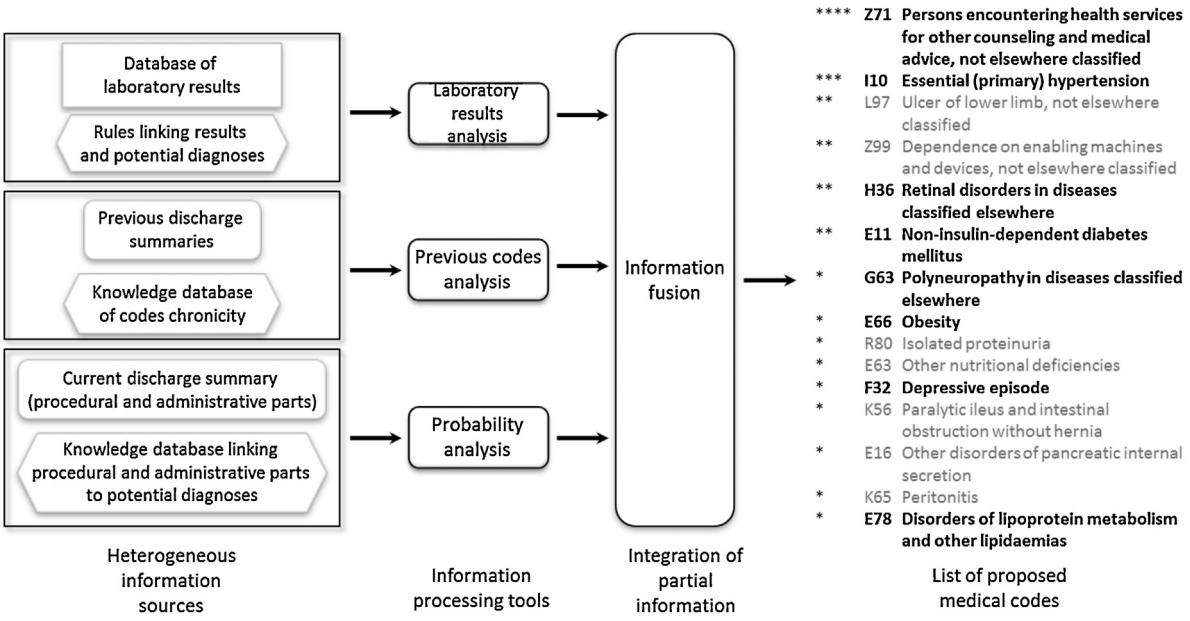

Fig. 2 – Main components of the used medical encoding support system, and example of a generated ranked codes list of an inpatient stay at the endocrinology clinical department. Symbols '*' indicate the evaluated degree of pertinence, and codes in bold were chosen by the physician.

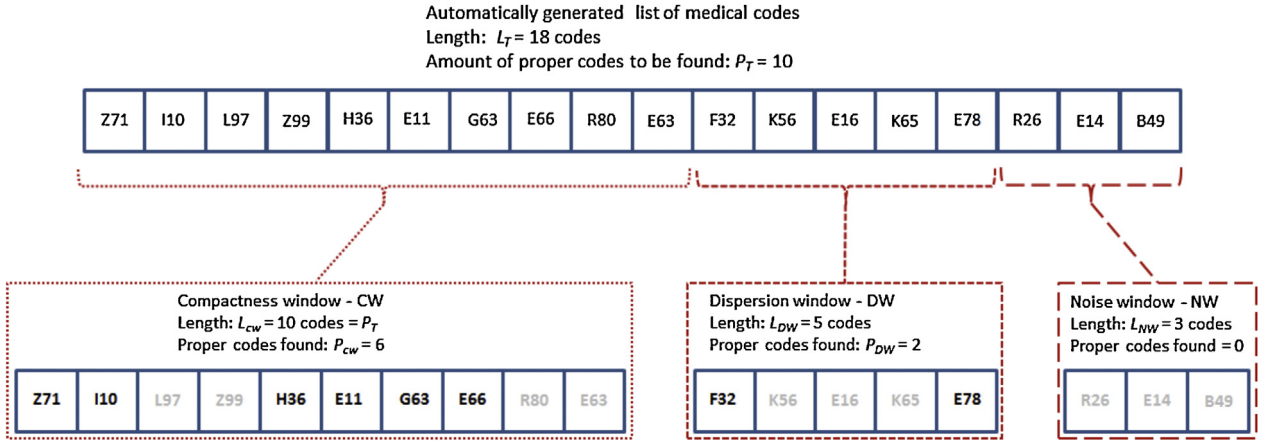

Fig. 3 – Structure of a codes list based on the example of Fig. 2, on which the physician expected to identify 10 proper codes and 8 were found (selected codes appear in bold).

2.2. *Usability study of the encoding support system*

Expert coders and physicians first review the available inpatient stay documents. Next, they study the corresponding codes list generated by the encoding support system, knowing from the documentation review how many diagnoses and procedures should be coded. Since the proposed codes list is formed by correct and incorrect codes, a strategy is applied to exploit it. The whole list is implicitly divided in three observation windows of varied lengths. Required correct codes should ideally be found in the first window, but if some or all expected correct codes are not in that window, the second window is inspected. The third window is partially or fully examined when correct codes are obviously still missing after the second window inspection. These three observation windows are called respectively, compactness, dispersion, and noise windows in the ADN quality measure model. Details of the coders' usability strategy are given in Fig. 3 and Table 1, indicating how coders' explanations were modeled in the ADN measure.

Conforming to the usability study, observation windows and their contents follow an implicit hierarchical structure. The first window (CW) is the most convenient, and is not affected by either dispersion or noise. Conversely, the second observation window (DW) is affected by dispersion and determines the length of the third window (NW), which begins after the last proper code found. Finally, when dispersion and noise are evaluated respectively in the second and third windows, resulting values add up to give a negative component that penalizes quality after the CW. These elements suggest that the global encoding information quality based on usability (IQ_L) has two distinct complementary components: quality in the compactness window (IQ_1), and in a combination of the dispersion and noise windows (IQ_2).

2.3. *Proposed ADN quality model for medical encoding information*

The information quality measure of a complete codes list was formalized interpreting the previous usability study, as expressed in the second column of Table 1 (ADN quality measure modeling).

2.3.1. *Quality component of the compactness window*

The number of correct codes (P_{CW}) found in the CW window defines the first component of the encoding information quality measurement (IQ_1). This accuracy value (A_{CW}), is defined as the ratio of proper codes to the total amount of necessary correct codes (P_T):

$$IQ_1 = A_{CW} = \frac{P_{CW}}{P_T} \tag{4}$$

Neither expert coders nor physicians consider the existence of incorrect codes in this first window disadvantageous for the exploitation of encoding support, because its full inspection is obvious.

2.3.2. *Quality components of the dispersion and noise windows*

In an equivalent manner, appropriate codes identified in the DW (P_{DW}) define the A_{DW} accuracy value, with respect to the total amount of necessary correct codes (P_T), as:

$$A_{DW} = \frac{P_{DW}}{P_T} \tag{5}$$

Contrarily to IQ_1, the second component of the encoding information quality measurement (IQ_2) is altered by dispersion and noise according to usability, and can be defined by:

$$IQ_2 = \underbrace{(A_{DW} - D)}_{DW} - \underbrace{N}_{NW} \tag{6}$$

where A_{DW} is the accuracy value of the DW observation window, D is the dispersion in that window, and N the noise in the NW window.

In order to represent the impact on information quality of dispersion and noise, two separate models were defined. Although on a practical basis the DW is examined before the NW, the next two sections describe first the noise and then the

Table 1 – Usability framework (abbreviations are defined in Fig. 3).

Usability description	ADN quality measure modeling
The total amount of expected proper codes to be identified is variable from one stay to another	P_T is a variable quantity
Depending on medical procedures and diagnoses the physician evaluates how many codes are required to code a given inpatient stay	The model parameter P_T is defined by the coder for the current inpatient stay
Any list generated by the encoding support system is expected to contain a number of proposed codes, greater or equal to the total of necessary proper codes to code a given inpatient stay	Expected length of a generated list $L_T \geq P_T$
A codes list is examined from left to right (or top to bottom) to look for correct codes, initially expecting to find all of them in the first places, and if that is not the case, in the rest of the list	The complete list is structured in three search windows, or sub-lists, characterized by quality factors: first, compactness window (CW), followed by dispersion window (DW), and noise window (NW)
The list of codes begins with the compactness window in which it is straightforward for the physician to look for the all the necessary codes; this window is considered as the most convenient	In the favorable case on which $L_T \geq P_T$: the length of the CW is $L_{CW} = P_T$
The presence of only part of the necessary correct codes in the compactness window, is not considered by physicians as being penalized by either dispersion or noise	Dispersion and noise in the CW are equal to zero; quality in CW is a positive value
If all expected proper codes are not found in the first part of the list, remaining proper codes can be found in the rest of it, except when the list length is inferior or equal to the amount of necessary proper codes	If $L_T \leq P_T$ then: the DW and NW cannot be defined, else: proper codes remaining to be found will be searched in the DW
The amount of proper codes expected to be found in the dispersion window, is at maximum equal to the difference between the total amount of necessary proper codes and the total amount of proper codes found in the compactness window	The number of expected correct codes P_{DW} in the DW, depends on the amount of proper codes P_{CW} found in the CW, so: $P_{DW} \leq P_T - P_{CW}$; and raw quality in DW is a positive value
The dispersion window length is defined by the last correct code found after the end of the compactness window	The length L_{DW} of the DW is variable; dispersion has a negative effect on quality
Codes remaining in the list after the dispersion window are incorrect and considered as noise, because the physician will look for additional correct codes without success, using more time unnecessarily	The length L_{NW} of the NW is variable; noise has a negative effect on quality

dispersion models, because the second one is mathematically based on the first.

2.3.2.1. Noise model. The subtraction term N in Eq. (6) acts as a penalization variable to information quality. Noise is produced by one or more incorrect codes positioned after the last correct code found at the end of the DW. It represents the user's unsuccessful search for omitted correct codes, from the end of the DW to the end of the list. Several constraints and specification elements, resulting from the usability analysis, were taken into account to build the noise model:

i. N must be an increasing function of the length L_{NW} of the NW.
ii. Quality reduction in the NW is proportional to the distance to the beginning of the window.
iii. N must be bounded for any length of the NW.
iv. N must increase proportionally to the amount of remaining codes expected to be found.

To satisfy condition (ii), the individual effect of an incorrect code (with index k in NW) is modeled as the value $f(x_k)$ of an increasing function f. The variable $x_k = \alpha k, \alpha \in \mathbb{R}^+$, models the code position on the support interval $[0,u]$, $u \in \mathbb{R}^+$, of f, associated to the whole NW. $f(x)$ can be defined as:

$$f(x) = ae^{bx} \tag{7}$$

To fulfill requirement (i), the noise penalization is defined as the sum of contributions assigned by f to each incorrect code in the NW. Determining parameters b and u was thus carried out by dealing with requirement (iii) through normalization conditions:

$$f(u) = 1, \quad I_N = \int_0^u ae^{bx}dx = 1 \tag{8}$$

Given the analytical constraints applied to f, both parameters a and b are defined within [0,1]. The resulting shape of function f makes individual additions to noise relatively small at the beginning of the NW, and very significant for incorrect codes located afar in the NW.

Summing individual contributions is equivalent to perform a numerical approximation of I_N in Eq. (8), ensuring convergence to 1 when L_{NW} increases. This sum of individual contributions is finally weighted by a value $\eta \in \mathbb{R}$, namely the noise factor, representing the remaining proportion of correct codes to be found after the DW. The noise factor value was set to $\eta = 1 - A_{CW} - A_{DW}$, satisfying the proportionality relationship of condition (iv). N is determined as:

$$N = \eta \cdot \sum_{k=1}^{L_{NW}} f(x_k) \cdot \Delta_N \tag{9}$$

where $\Delta_N = x_{k+1} - x_k = u/L_{NW}$ is the width of rectangles in the numerical integration formula.

2.3.2.2. Dispersion model. Contrary to the NW, the DW may contain a set of correct codes that corresponds to all or only a part of the remaining proper codes to be found after the end of the first observation window. However, if the DW also contains incorrect codes, dispersion of proper codes reduces the quality calculated in Eq. (5). For that reason, the term D in Eq. (6) penalizes directly the second information quality component depending on proper codes scattering in the DW. Dispersion of proper codes in DW was estimated, applying a weighting function g, defined on a support interval $[0,v]$, $v \in \mathbb{R}^+$, under equivalent mathematical constraints assumed to characterize noise. In particular, codes contribution to quality must diminish as the search point moves away from the beginning of the DW. $g(x)$ is represented in the model by:

$$g(x) = 1 - ce^{dx} \tag{10}$$

Parameter c was also chosen within [0,1] and the value d deduced from analytical constraints on g.

The total information quality of the DW, defined by term $(A_{DW} - D)$ in Eq. (6), should decrease to 0 when dispersion increases significantly, and according to the size of DW. Given that the sum of individual codes contribution in D is constrained inside [0,1], a dispersion factor $\delta \in \mathbb{R}$, equal to the ratio A_{DW} was applied. As a result, when the number of incorrect codes in the DW increases, the sum of individual contributions tends to 1 and D converges to $\delta = A_{DW}$, as inferred from the usability study. Hence, D is calculated as:

$$D = \delta \cdot \sum_{k=1}^{L_{DW}} \phi_k \cdot \{g(x_k)\Delta_D\} \tag{11}$$

where the binary coefficient $\phi_k = 1$ if the code at position k is incorrect, $\phi_k = 0$ otherwise (to account only the presence of incorrect codes), and $\Delta_D = x_{k+1} - x_k = v/L_{DW}$ can be different than Δ_N in Eq. (9).

2.3.3. *Representation of the ADN information quality measurement model*

Elements IQ_1 and IQ_2 are two distinct but complementary components of the quality measurement, suggesting that a 2D point should depict their respective contribution at a representation level. According to Eq. (4), information quality IQ_1 varies in the interval [0,1]. Likewise, Eq. (6) specifies information quality IQ_2 is within [−1,1], due to the combined influence of noise and dispersion. Usability indicates that it is possible to have a global quality $IQ_L \neq 0$ even when one of the quality components, either IQ_1 or IQ_2, is zero. For that reason, the combination of IQ_1 and IQ_2 cannot be multiplicative, and clearly appears to be additive. Therefore, the global quality model (IQ_L) can be built adding IQ_1 and IQ_2 as follows:

$$IQ_L = IQ_1 + IQ_2 = A_{CW} + (A_{DW} - D) - N \tag{12}$$

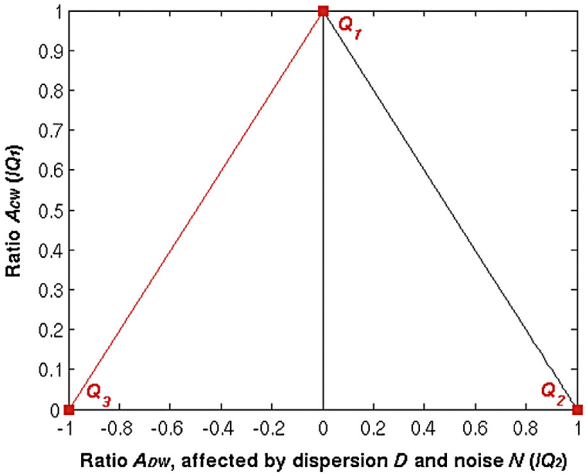

Fig. 4 – Representation space of the ADN information quality measure model.

$$IQ_L = \frac{P_{CW}}{P_T} + \left(\frac{P_{DW}}{P_T} - \delta \cdot \sum_{k=1}^{L_{DW}} \phi_k \cdot \{g(x_k) \cdot \Delta_D\} \right) - \sum_{k=1}^{L_{NW}} f(x_k) \cdot \Delta_N \tag{13}$$

Usability implies that the user should be able to simultaneously quantify the global quality and identify each component IQ_1 and IQ_2. Both conditions are satisfied when drawing quality points associated to values IQ_L in the 2D representation space associated to the valid intervals of each quality component. For ease of visualization, it was decided to assign the second component (IQ_2 in [−1,1]) to abscises, and the first one (IQ_1 in [0,1]) to ordinates. Due to constraints resulting from models of quality components, the representation space is limited by three numerical boundaries that define a triangular space (Fig. 4):

- If $y = IQ_1 = 1$, then $IQ_2 = 0$ with $D + N = 0$, hence $x = 0$, defining point $Q_1 = (0, 1)$.
- If $y = IQ_1 = 0$ and $IQ_2 = 1$, then $A_{DW} = 1$ and $D + N = 0$, defining point $Q_2 = (1, 0)$.
- If $y = IQ_1 = 0$ and $IQ_2 = -1$, then $A_{DW} = 0$, $D = 0$ and $N = -1$, defining point $Q_3 = (-1, 0)$.

The best information quality is found at Q_1, when all expected correct codes are identified in the CW. This value is equal to producing the same list as the expert coder. Equivalent results are to find all expected codes distributed between the CW and the beginning of the DW, or only at the beginning of the DW (point Q_2). Nevertheless, if for instance all expected correct codes are found in the generated list, with variable combinations of dispersion and noise, different values of IQ_L inferior to 1 result. All these values are properly represented in the defined measurement space. Furthermore, the worst quality measure is the point Q_3 when the list only contains incorrect codes.

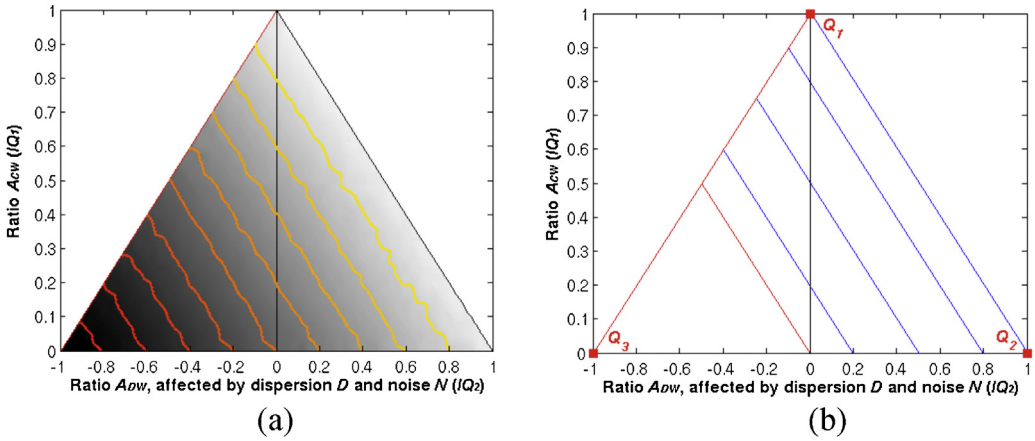

Fig. 5 – Quality measurement diagram: (a) quality values of 10,000 simulated codes list and examples of iso-quality lines; (b) pre-defined reference iso-quality lines.

3. Results

The ADN information quality measure was tested using simulated and real data sets in order to verify the theoretical model and show its usefulness, respectively.

3.1. *Information quality of simulated medical codes lists*

Correct and incorrect codes were selected randomly from a reference database, making parameters change independently. Except for the next sub-section, simulations correspond to average values of 100 generated lists, representing each particular test context.

3.1.1. *Validation of the quality representation spatial boundaries*

Quality measurements of 10,000 simulated codes lists were visualized on the same diagram. Results clearly show that the ADN measures representation fitted the inferred triangular space, defined by Q_1, Q_2, and Q_3 in Fig. 4. Obtained information quality values covered the range of possible measurements. When represented proportionally to gray levels, these values made emerge iso-quality lines (slightly irregular in Fig. 5(a) due to simulation discontinuities). These iso-quality lines appear oriented in parallel to the line that connects the maximum quality points (Q_1 and Q_2).

The isosceles triangle is defined on the right by the maximum iso-quality line, which only represents accuracy components. The line of maximum noise effect defines the left edge of the triangle. This boundary line connects the points of maximum (Q_1) and minimum (Q_3) information quality. Among countless possibilities, five examples of quality zones were defined in Fig. 5(b), delimited by iso-quality lines at 0.8, 0.5, 0.2, and 0. From right to left, these five arbitrarily defined zones represent decreasing information quality, with gradually augmented impact of dispersion and noise. Below iso-quality line 0, dispersion and noise deteriorate completely the accuracy of the DW.

3.1.2. *Variable size of the dispersion and noise windows*

The CW size was defined by parameter P_T, and the size of the DW by the position of the last correct code identified in the list. The simulations were carried out supposing that out of 12 expected correct codes, $P_{CW} = 1$ was found in the CW to emphasize the effect of dispersion, and $P_{DW} = 10$ in the DW, without noise (Fig. 6(a)). L_{DW} size was modified gradually augmenting dispersion, by adding up to 100 incorrect codes.

Dispersion simulation started with value 0 and converged to the dispersion factor $\delta = A_{DW}$, corresponding to the proportion of proper codes found in the DW. As a consequence, quality in the DW tends to 0, confirming the expected theoretical behavior. Augmenting dispersion by decreasing correct codes in the DW, keeping the window size constant, has the same effect on quality.

The impact of noise was evaluated changing the length of the NW from 0 to 100 incorrect codes (Fig. 6(b)). To underline the representation of noise, the quantities of correct codes found in the CW and the DW were 1 in both cases, among 12 expected codes, without dispersion. Values evolution was again consistent with the theoretical assumptions, since noise measurement started from 0 and converged to $\eta = 1 - A_{CW} - A_{DW}$. Quality measurement points were displaced to the left of the diagram describing a horizontal trajectory, following the increase of improper codes in the NW.

3.2. *Compared information quality of simulated lists*

To illustrate simple cases of comparative evaluation, the ADN information quality measure was applied to the simulated lists of the six examples illustrated in Fig. 1, for which conventional precision-recall based approaches do not respond to fitness for use questions. Results in Fig. 7 show that: case 2 has better information quality than case 1 although both quality values are in the same zone; case 4 has lower quality than case 3 as a consequence of dispersion and noise; case 5 has lower quality than case 6 for the same reason.

Additionally, two sets of ten simulated codes lists, over separate quality zones, were also evaluated comparatively with respect to a precision-recall analysis. The ADN quality

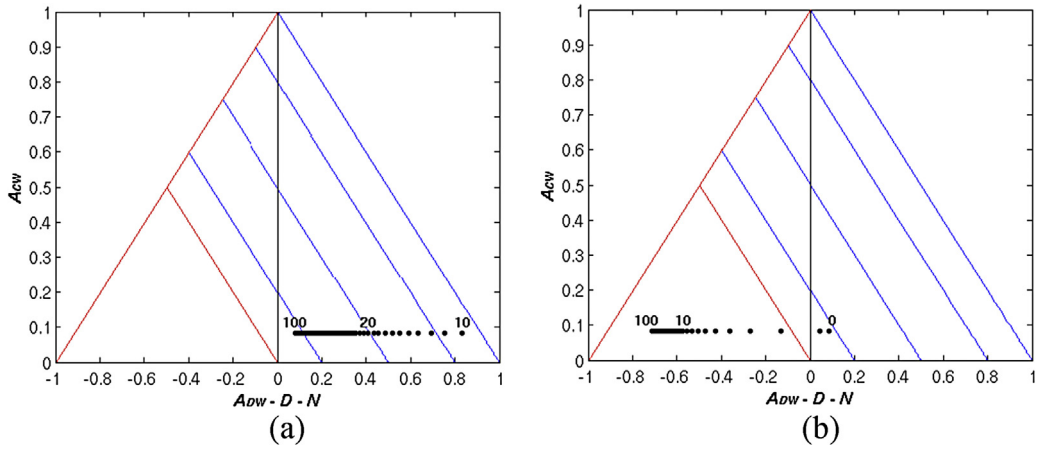

Fig. 6 – Effect of variable size windows on: (a) dispersion and (b) noise (points are identified by the variable length of DW or NW respectively).

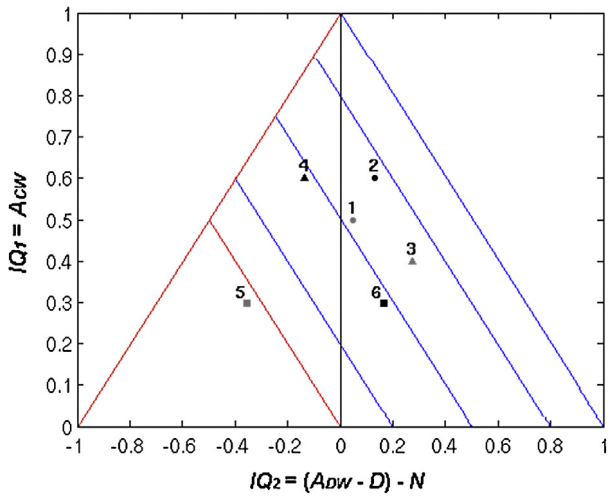

Fig. 7 – Compared information quality evaluation by pairs using the ADN model, of the six lists represented as gray and dashed black lines using precision-recall in Fig. 1: cases 1 and 2 (points); cases 3 and 4 (triangles); cases 5 and 6 (squares).

measure of the 20 codes lists (Fig. 8(a)) provides a clear individual usability evaluation, even when quality values are in neighbor zones. Conversely, the two precision-recall diagrams (Fig. 8(b and c)) cannot be clearly analyzed giving the superposition of lines, and the lack of other evaluation criteria. Moreover, the variable behavior of each curve impedes to compare curves sets when recall is less than 0.5. Using the ADN measure within a quality zone appears more suitable to represent and compare sets of quality values.

3.3. *Information quality of real medical encoding support lists*

Since coded inpatient stays were randomly selected, lengths of stay, quantity of diagnoses per stay, and types of diagnoses codes were not uniformly distributed. Nevertheless, this set of codes lists was considered by expert coders as representative of each clinic department activity. Additionally, experts' information quality assessment corresponds to the best possible, given that it is equivalent to find all the expected codes in the first observation window (CW). In order to determine the degree of ranking coherence between the ADN measure and five commonly used information retrieval measures, the Spearman's rank coherence coefficient [49] was calculated between order indexes of codes lists by pairs,

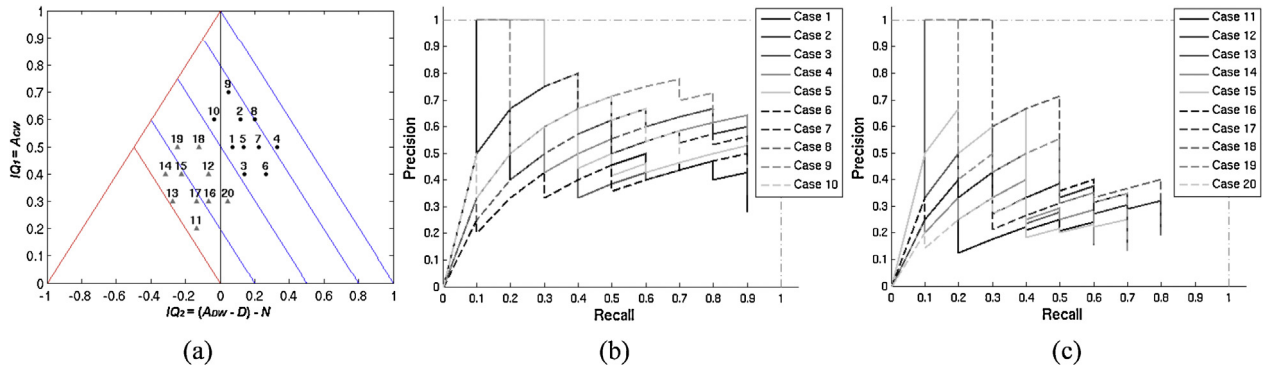

Fig. 8 – Compared information quality representation of twenty different codes lists, using the ADN model (a), and applying precision-recall (b and c).

Table 2 – Compared Spearman's rank coherence coefficients between list orders obtained with ADN and five information retrieval measures, corresponding to four clinic departments.

	Traumatology	Obstetrics	Urology	Cardiology
Mean precision	0.55	0.59	0.53	0.7
Mean recall	0.55	0.64	0.76	0.69
Final precision	0.67	0.61	0.45	0.62
Final recall	0.61	0.72	0.73	0.71
Mean Fm	0.72	0.71	0.61	0.73

evaluated by ADN measure, and by five commonly used information retrieval measures. This coefficient does not depend on the assumption of a given underlying statistical distribution. In our case, the Spearman's rank coherence coefficient determines to which extent there are pairwise agreements, between a ranked set of ADN measures and a ranked set of information retrieval measure values. These information retrieval measures are the: mean precision and recall along *Pr–Rc* curves, precision and recall at the end of *Pr–Rc* curves (final precision and final recall), and mean *Fm*-measure values (Table 2). The codes lists for this test were automatically generated by the coding support system (Section 2.1).

Results in Table 2 are categorized according to three segments determined from the values distribution in the interval [0,1], being the most important in our case: [0.45, 0.55), [0.55, 0.65) and [0.65, 0.75]. These values show a variable positive compared ranking coherence, but the ADN measure is not perfectly monotonically related to the information retrieval measures, in agreement with its expected behavior. The rows of Table 2 show that ADN is globally somehow more compatible with recall than with precision. This can be explained considering that the measure has been built to be especially severe with wrong codes (dispersion and noise), in accordance with the usability study. Additionally, it is likely to have a more classical behavior with respect to missing codes. The ADN measure is nevertheless relatively consistent with the F-measure, indicating its capability to express quality as a compromise between the correctness and the completeness of a codes list. It is also interesting to note that the coherence result varies depending on the clinic department. These results show that the ADN measure represents information quality in an alternative consistent way. However, the particular differences with respect to other measures cannot be meaningfully explained for each clinic department by the compared Spearman's coefficient.

To illustrate the usefulness of our approach, lists of plausible codes automatically generated by the coding support system described in Section 2.1 were processed applying the ADN measure. Values were visualized using the ADN triangular quality diagram, and compared to an equivalent information retrieval diagram based on classical measures. Fig. 9 depicts the results for the four clinic departments. Each dot represents the information quality of one complete codes list evaluated with the ADN measure (Fig. 9(a,c,e and g)) and the corresponding final precision-recall value (Fig. 9(b,d,f and h)). Level curves of points' concentration on the plane were also defined to identify the main information quality trends in a simple manner. This enables distinguishing regions with isolated points (light-gray/white) from regions with clusters of superposed points (dark-gray).

Results indicate that the ADN information quality of codes lists generated by the encoding support system is spread to some extent, depending on the clinic department. Information quality points for a significant amount of the generated codes lists appear on the left side of the diagram, due to the presence of dispersion and noise. Nevertheless, depending on the clinic department, dots concentrated on different quality zones, providing a global evaluation of quality measurement trends by clinic department. This illustrates the fact that inputs to the encoding system were comparatively heterogeneous for each clinic department.

In the case of the traumatology department (Fig. 9(a)), a considerable amount of information quality measurements appeared below the iso-quality line 0.2, being the main cluster below 0. In agreement with usability, such quality values are not exploitable for encoding support, given their low levels. For the obstetrics department, most points are located between the 0 and 0.5 iso-quality lines, and a set of points above 0.5 (Fig. 9(c)), with reduced dispersion and noise if compared to traumatology and urology. Information quality of the urology department codes lists (Fig. 9(e)), concentrated between the 0.2 and 0.5 iso-quality lines, having a secondary cluster between 0 and 0.2, but with more codes in the CW than for other clinic departments. Information quality of the lists produced for the cardiology department (Fig. 9(g)), also concentrated in a main cluster between the 0.2 and 0.5 iso-quality lines, showing although the most reduced impact of dispersion and noise compared to other clinic departments. These results show how the notions of codes dispersion and noise are appropriately represented by the ADN quality measure. From a global perspective, lists of the obstetrics department can be considered as of intermediate quality compared to urology and cardiology. Lists from the cardiology and urology departments appear over the same iso-quality zones, but with different characteristics. Results for the urology department could be considered as more reliable than for the cardiology department, mainly because of the superior proportions of correct codes in the CW. Nevertheless, evaluating overall efficiency, results for the cardiology department could be faster to process by human coders given the reduced dispersion and noise.

Final precision-recall representations confirm all previous indications of information retrieval performance measures limitations to evaluate coding lists quality based on usability. Considering the four representations (Fig. 9(b,d,f and h)), it is unclear how to identify coherent trends in the sense of fitness for use, except an appreciation of the points sets position to the left of the right of the diagram. Measure points in these

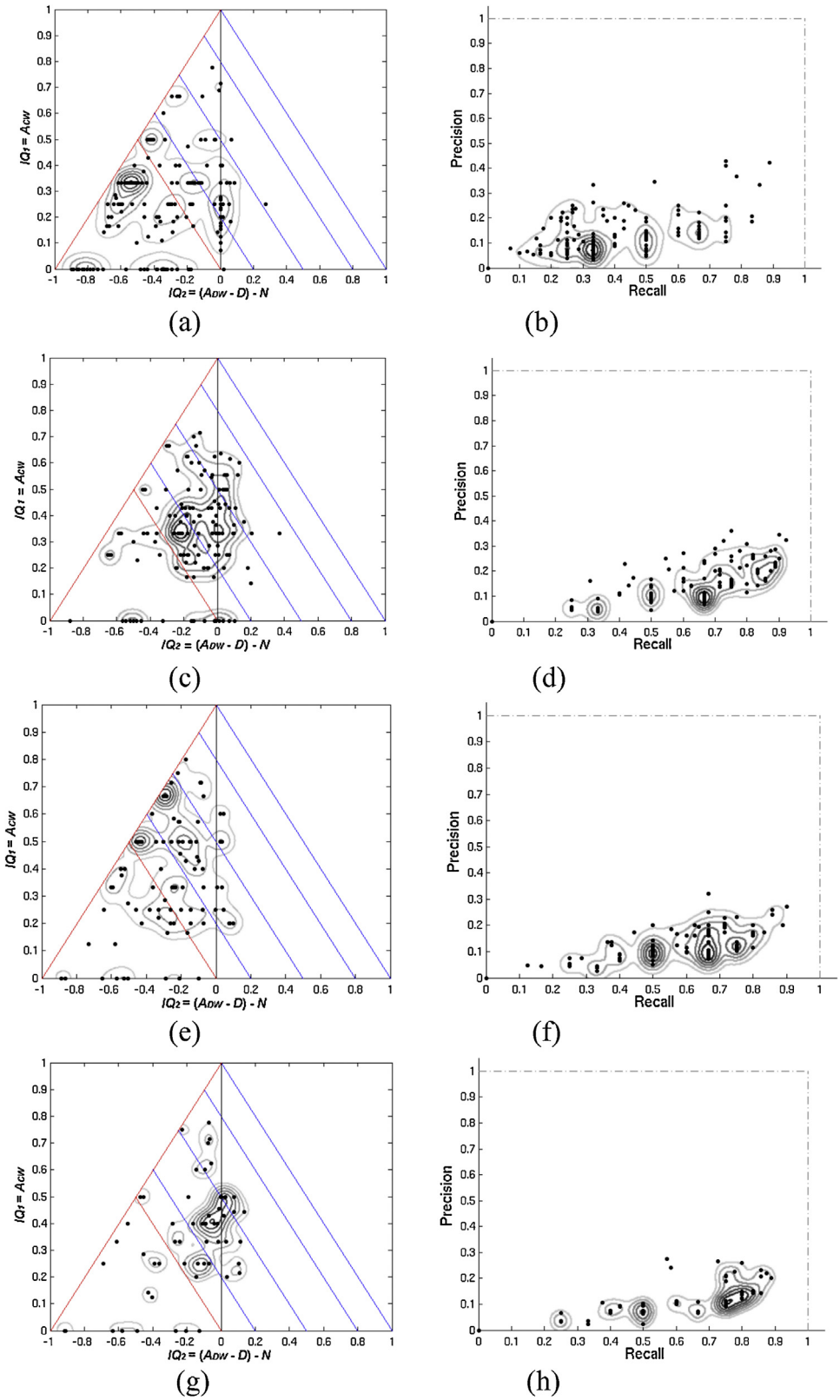

Fig. 9 – Information quality points of codes lists generated by the encoding-support system and corresponding density-level curves, for the traumatology (a,b) obstetrics (c,d) urology (e,f) and cardiology (g,h) clinic departments of the same hospital, calculated with the ADN (a,c,e,g) and the final precision-recall (b,d,f,h) measures.

figures are highly concentrated without a structure of quality components dimensions and zones. Additionally, codes list with equal amounts of correct and incorrect codes but distributed differently, also appear as superposed or very close points in the diagrams. These results provide complementary indices about the suitability of the ADN measure to represent information quality of medical coding support lists, based on their usability.

4. Discussion

Conformingly to the usability study, the ADN information quality measure was designed to match, as close as possible, the stated facts and understanding. The ADN measure components are determined in three observation windows of size adapted to analyzed data, instead of one fix length window applied by information retrieval measures. Otherwise, the ADN quality measure properly represents two components of information quality evaluation, as a unique point in a constrained numerical space, rather than a precision-recall curve. A significant finding of this quality model is the definition of iso-quality lines, which represent all the points having the same IQ_L value, with different quality components contributions.

This representation is quite revealing in several ways. First, it depicts clearly the fact that a given quality value can result from countless combinations of the quality factors. Second, unlike 2D accuracy diagrams, the quality variation between two lists with the same correct codes, but organized differently, can be detected in a simple manner along the iso-quality lines. Third, these iso-quality lines define flexible information quality zones. The interpretation of the diagram iso-quality lines is of particular interest. When two different points of the same iso-quality line were proposed, the one with more proper codes in the CW is preferred by the users. That point is placed at a higher position on the iso-quality line, although global quality values are equal. The diagram could be further detailed depending on how encoding information quality evaluation is carried out, for instance narrowing and incrementing quality zones, and/or subdividing the existing zones. It is also noteworthy to observe that the information quality of hundreds of codes lists can be easily compared using only one diagram.

Analysis of the clusters formed by measured information quality points enables to perform an assessment of encoding support system efficiency. An implication of these results is the possible conception of complementary representations, according to other usability factors. Even if for this study the four information quality terms of Eq. (12) were analyzed, some others could be integrated. For example: an estimation of the maximum number of examined codes, dispersion and noise in the first observation window, the added dispersion of the first two observation windows, or noise-dispersion ratios. As a consequence, additional usability dependent quality dimensions can be examined accordingly.

Functions used in the models characterize inferred constraints associated to the independent progressions of dispersion and noise. Therefore the impact of each correct or incorrect code to information quality is considered within an observation window. Both dispersion and noise are interpreted as different degrees of uncertainty that have variable impact on the encoding support usefulness of a list. Dispersion and noise produce cumulative information quality diminution, interpreted as factual for dispersion and conditional for noise. In the case of noise, the true value will depend on the length of the list that the expert coder or physician decides to inspect. At a practical level, a human coder will certainly experience the hesitation and delay produced by dispersion, but not necessarily the estimated noise setback. When the coder thinks that it is not worth to look for more correct codes, the inspection can be halted at any point in the list. Estimated noise represents nevertheless how severe could be information quality loss if the coder examines the third observation window completely. Not knowing how far each coder will go on practical conditions, an estimated value could be added to the model, proportionally to the amount of missing correct codes. For instance it should be relatively weak when nearly all codes are identified in the CW, and stronger when most of the correct codes were not found in the DW.

Simulation tests demonstrated that the ADN information quality model behaved coherently within the bounds of the triangular representation diagram, according to the theoretical definition. Conversely, results of the validation with real data suggest that the noise penalization model may be severe, given that it displaced to the left of the triangular diagram a considerable quantity of points. Some modifications could be envisaged, like the application of weighting functions to adjust the proposed representation of dispersion and noise, with adapted bounding conditions.

5. Conclusion

Medical encoding support is an emerging technology based on complementarities of information sources, for which adapted information quality measures are required, independently of the sources heterogeneous nature. This work set out to determine a usability-based information quality measure of medical encoding support lists. A study of encoding support systems usability was analyzed to define, characterize, and validate a relevant encoding information quality model. Usability relayed on the coder strategy to find the expected amount of correct codes in the generated list, using three observation windows. Four usability factors compose the defined information quality measure: proportions of correct codes found in the first and second observation windows, dispersion of correct codes in the second window, and noise resulting from incorrect codes in the third window. The information quality measure was consistently bounded to enable comparing codes lists of different lengths, containing variable quantities and distributions of correct and incorrect codes. This model enhances the understanding of medical encoding information quality measurement and considerably improves analysis based on 2D precision-recall curves. The most significant findings to emerge from this study are the information quality components and the original triangular space representation of the information quality measure. All these elements enable to analyze medical encoding support systems reliability, for instance considering the clusters of quality points associated to different quality zones. On the other

hand, this work only examined a type of functions for dispersion and noise. Several other functions could also be examined for that purpose, even making use of a different type of function for each factor. Further research will be undertaken to investigate automated comparative indexes to dynamically adapt costs of penalizations, and include advanced clustering approaches to identify encoding performance patterns. Moreover, the fitness for use is constrained to medical encoding support systems. For this reason its applicability to other information quality problems would require the identification of distinctive usability factors, associated to the given particular application.

Acknowledgement

This work was supported by a TECSAN/ANR project under the name Med1Dex associating the university hospital CHRU Brest – France, Telecom Bretagne, INSERM Unit U1101 LaTIM, and PRISMEDICA. Authors are very grateful to D. Gueriot for his enriching comments and to the anonymous reviewers for their constructive suggestions.

REFERENCES

[1] P. Cheng, A. Gilchrist, K.M. Robinson, L. Paul, The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding, Health Information Management Journal 38 (1) (2009) 35–46.
[2] J.J. Cimio, Coding systems in health care, Methods of Information in Medicine 35 (4–5) (1996) 273–284.
[3] C.J. Buck, J.K. Grass, Step-by-Step Medical Coding, 2011 ed., Elsevier, Saunders, 2011.
[4] S.A.R. Nouraei, S. O'Hanlon, C.R. Butler, A. Hadovsky, E. Donald, E. Benjamin, G.S. Sandhu, A multidisciplinary audit of clinical coding accuracy in otolaryngology: financial, managerial and clinical governance considerations under payment-by-results, Clinical Otolaryngology 34 (1) (2009) 43–51.
[5] W.C. Morris, D.T. Heinze, H.R. Warner, A. Primack, A.E. Morsch, R.E. Sheffer, M.A. Jennings, M.L. Morsch, M.A. Jimmink, Assessing the accuracy of an automated coding system in emergency medicine, in: Proceedings of the Symposium of the American Medical Informatics Association, 2000, pp. 595–599.
[6] D.T. Heinze, M.L. Morsch, R.E. Sheffer, M.A. Jimmink, M.A. Jennings, W.C. Morris, A.E.W. Morsch, LIFECODE: a deployed application for automated medical coding, Artificial Intelligence Magazine 22 (2) (2001) 76–88.
[7] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, Journal of the American Medical Informatics Association 11 (2004) 392–402.
[8] S.V.S. Pakhomov, J.D. Buntrock, C.G. Chute, Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques, Journal of the American Medical Informatics Association 13 (5) (2006) 516–525.
[9] R. Farkas, G. Szarvas, Automatic construction of rule-based ICD-9-CM coding systems, BMC Bioinformatics 9 (Suppl. 3) (2008) S10 http://www.biomedcentral.com/1471-2105/9/S3/S10
[10] M. Hubert, C. Le Guillou, F. Le Saux, L. Lecornu, J.-M. Cauvin, iRMA: web interface for ICD10 code and context retrieval, in: Proceedings of the Patient Classification Systems International, 2010, pp. 53–54.
[11] T. Imai, M. Kajino, M. Sato, K. Ohe, Development of structured ICD-10 and its application to computer-assisted ICD coding, Studies in Health Technology and Informatics 160 (2) (2010) 1080–1084.
[12] S. De Lusignan, C. Minmagh, J. Kennedy, M. Zeimet, H. Bommezijn, J. Bryant, A survey to identify the clinical coding and classification systems currently in use across Europe, in studies in health technology and informatics, in: Proceedings of the MEDINFO, 2001, pp. 86–89.
[13] K. McKenzie, S.M. Walker, C. Dixon-Lee, G. Dear, J. Moran-Fuke, Clinical coding internationally: a comparison of the coding workforce in Australia, America, Canada and England, in: Proceedings of the 14th International Federation of Health Records Congress, 2004 http://library.ahima.org/xpedio/groups/public/documents/ahima/bok3_005534.hcsp?dDocName=bok3_005534
[14] S. Santos, G. Murphy, K. Baxter, K.M. Robinson, Organizational factors affecting the quality of hospital clinical coding, Health Information Management Journal 37 (1) (2008) 25–37.
[15] Australian Institute of Health and Welfare, The coding Workforce Shortfall, Catalogue No. HWL 46, AIHW, Canberra, 2010.
[16] D.M. Strong, Y.W. Lee, R.Y. Wang, Data quality in context, Communications of the ACM 40 (5) (1997) 103–110.
[17] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, D. Lee, A taxonomy of dirty data, Data Mining and Knowledge Discovery 7 (1) (2003) 81–99.
[18] R.Y. Wang, D.M. Strong, Beyond accuracy: what quality means to data consumers, Journal of Management Information Systems 12 (4) (1996) 5–34.
[19] D.G.T. Arts, N.F. De Keizer, G.-J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework, Journal of the American Medical Informatics Association 9 (2002) 600–611.
[20] L.L. Pipino, Y.W. Lee, R. Wang, Data quality assessment, Communications of the ACM 45 (4) (2002) 211–218.
[21] Y.W. Lee, D.M. Strong, B.K. Khan, R.Y. Wang, AIMQ: a methodology for information quality assessment, Information and Management 40 (2002) 133–146.
[22] W.E. Winkler, Methods for evaluating and creating data quality, Information Systems 29 (7) (2004) 531–550.
[23] M.H. Stanfill, M. Williams, S.H. Fenton, R.A. Jenders, W.R. Hersh, A systematic literature review of automated clinical coding and classification systems, Journal of the American Medical Informatics Association 17 (2010) 646–651.
[24] B. Stvilia, L. Gasser, M.B. Twidale, L.C. Smith, A framework for information quality assessment, Journal of the American Society for Information Science and Technology 58 (12) (2007) 1720–1733.
[25] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Computing Surveys 41 (3) (2009), Article 16.
[26] A. Jamal, K. McKenzie, M. Clark, The impact of health information technology on the quality of medical and health care: a systematic review, Health Information Management Journal 38 (3) (2009) 26–37.
[27] C.J. van Rijsbergen, Information Retrieval, 2nd ed., Butterworth Scientific Ltd, Surrey, UK, 1979.
[28] V. Raghavan, P. Bollmann, G.S. Jung, A critical investigation of recall and precision as measures of retrieval system performance, ACM Transactions on Information Systems 7 (3) (1989) 205–229.

[29] J. Davis, M. Goadrich, The relationship between precision-pecall and ROC curves, in: Proceedings of the 23rd ACM International Conference on Machine Learning, 2006, pp. 233–240.
[30] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, Advances in Artificial Intelligence LNCS 4304 (2006) 1015–1021.
[31] D. Hull, Using statistical testing in the evaluation of retrieval experiments, in: Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information retrieval, 1993, pp. 329–338.
[32] T. Sakai, Ranking the NTCIR systems based on multigrade relevance, in: Proceedings of the International Conference on Asian Information Retrieval Technology, 2004, pp. 251–262.
[33] A. Moffat, J. Zobel, Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems 27 (1) (2008) 1–27, 2.
[34] C. Buckley, E.M. Voorhees, Retrieval evaluation with incomplete information, in: Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 25–32.
[35] W.R. Hogan, M.W. Wagner, Accuracy of data in computer-based patient records, Journal of the American Medical Informatics Association 5 (1997) 342–355.
[36] P.S. Romano, S. Yasmeen, M.E. Schembri, J.M. Keyzer, W.M. Gilbert, Coding of perineal lacerations and other complications of obstetric care in hospital discharge data, Obstetrics and Gynecology 106 (4) (2005) 717–725.
[37] P. Resnik, M. Niv, M. Nossal, G. Schnitzer, J. Stoner, A. Kapit, R. Toren, Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding, in Perspectives in Health Information Management, in: Proceedings of the Computer Assisted Coding Conference, 2006 http://www.library.ahima.org/xpedio/groups/public/documents/ahima/bok1_032010.html
[38] T.-H. Lu, M.-C. Lee, M.-C. Chou, Accuracy of cause-of-death coding in Taiwan: types of miscoding and effects on mortality statistics, International Journal of Epidemiology 29 (2000) 336–343.
[39] K. Jordan, M. Porcheret, P. Croft, Quality of morbidity coding in general practice computerized medical records: a systematic review, Family Practice 21 (4) (2004) 396–412.
[40] T. Henderson, J. Shepheard, V. Sundararajan, Quality of diagnosis and procedure coding in ICD-10 administrative data, Medical Care 44 (11) (2006) 1011–1019.
[41] R. Geierhofer, A. Holzinger, The evaluation of semantic tools to support physicians in the extraction of diagnosis codes, in: A. Holzinger (Ed.), HCI and Usability for Medicine and Health Care, LNCS 4799, Springer, Berlin, Heidelberg DE, 2007, pp. 403–408.
[42] D.T. Heinze, M.L. Morsch, B.C. Potter, R.E. Sheffer, Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology, Journal of the American Medical Informatics Association 15 (1) (2008) 40–43.
[43] M.K. Lam, K. Innes, P. Saad, J. Rust, V. Dimitropoulos, M. Cumerlato, An evaluation of the quality of obstetric morbidity coding using an objective assessment tool, the performance indicators for coding quality (PICQ), Health Information Management Journal 37 (2) (2008) 19–29.
[44] L. Lecornu, C. Le Guillou, G. Thillay, P.J. Garreau, H. Jantzem, J.-M. Cauvin, C2i: a tool to gather medical indexed information, in: Proceedings of the 9th IEEE International Conference on Information Technology and Applications in Biomedicine, 2009, http://dx.doi.org/10.1109/ITAB. 2009.5394351.
[45] L. Lecornu, C. Le Guillou, F. Le Saux, M. Hubert, J. Puentes, J.-M. Cauvin, ANTEROCOD: actuarial survival curves applied to medical coding support for chronic diseases, in: Proceedings of the 32nd IEEE International Conference of the Engineering in Medicine and Biology Society, 2010, pp. 1158–1161.
[46] L. Lecornu, G. Thillay, C. Le Guillou, P.J. Garreau, P. Saliou, H. Jantzem, J. Puentes, J.-M. Cauvin, REFEROCOD: a probabilistic method to medical coding support, in: Proceedings of the 31st IEEE International Conference of the Engineering in Medicine and Biology Society, 2009, pp. 3421–3424.
[47] L. Lecornu, C. Le Guillou, F. Le Saux, M. Hubert, J. Puentes, J. Montagner, J.-M. Cauvin, Information fusion for diagnosis coding support, in: Proceedings of the 33rd IEEE International Conference of the Engineering in Medicine and Biology Society, 2011, pp. 3176–3179.
[48] World Health Organization, International classification of diseases: ICD-10, vol. 1–3, 2007 http://www.apps.who.int/classifications/icd10/browse/2010/en
[49] R. Kumar, S. Vassilvitskii, Generalized distances between rankings, in: Proceedings of the 19th ACM International Conference on World Wide Web, 2010, pp. 571–580.