Ann Ig 2023 Jan-Feb; 35(1): 3-20 doi: 10.7416/ai.2022.2517. Epub 2022 Apr 12.

Natural language processing and String Metric-assisted Assessment of Semantic Heterogeneity method for capturing and standardizing unstructured nursing activities in a hospital setting: a retrospective study

M. Vanalli[1], M. Cesare[1], A. Cocchieri[2], F. D'Agostino[3]

Key words: Cross-mapping, natural language processing, standardized nursing terminology, professional assessment instrument, nursing activities, clinical nursing information system

Parole chiave: Cross-mapping, elaborazione del linguaggio naturale, terminologia infermieristica standard, professional assessment instrument, attività infermieristiche, sistema informativo infermieristico

Abstract

Background. *Nurses record data in electronic health records (EHRs) using different terminologies and coding systems. The purpose of this study was to identify unstructured free-text nursing activities recorded by nurses in EHRs with natural language processing (NLP) techniques and to map these nursing activities into standard nursing activities using the SMASH method.*

Study design. *A retrospective study using NLP techniques with a unidirectional mapping strategy called SMASH.*

Methods. *The unstructured free-text nursing activities recorded in the Medicine, Neurology and Gastroenterology inpatient units of the Agostino Gemelli IRCCS University Hospital Foundation, Rome, Italy were collected for 6 months in 2018. Data were analyzed by three phases: a) text summarization component with NLP techniques, b) a consensus analysis by four experts to detect the category of word stems, and c) cross-mapping with SMASH. The SMASH method calculated the string comparison, similarity and distance of words through the Levenshtein distance (LD), Jaro-Winker distance and the cross-mapping's cut-offs: map [0.80-1.00] with < 13 LD, partial-map [0.50-0.79] with <13 LD and no map [0.0-0.49] with >13 LD.*

[1] *Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy*

[2] *Fondazione Policlinico Universitario A. Gemelli IRCCS, University of Catholic Sacred Heart, Rome, Italy*

[3] *Unicamillus, Saint Camillus International University of Health Sciences, Rome, Italy*

Abbreviations. NLP: natural language processing; CNIS: clinical nursing information system; PAI: professional assessment instrument; SNT: standard nursing terminology; EHR: electronic health record

Annali di Igiene : Medicina Preventiva e di Comunità (Ann Ig)
ISSN 1120-9135 https://www.annali-igiene.it

Results. *During the study period, 491 patient records were assessed. 548 different unstructured free-text nursing activities were recorded by nurses. 451 unstructured free-text nursing activities (82.3%) were mapped to standard PAI nursing activities, 47 (8.7%) were partial mapped, while 50 (9.0%) were not mapped. This automated mapping yielded recall of 0.95%, precision of 0.94%, accuracy of 0.91%, F-measure of 0.96. The F-measure indicates good reliability of this automated procedure in cross-mapping.*

Conclusions. *Lexical similarities between unstructured free-text nursing activities and standard nursing activities were found, NLP with the SMASH method is a feasible approach to extract data related to nursing concepts that are not recorded through structured data entry.*

Introduction

Over the past few years, healthcare organizations have been working to implement visible and structured nursing data within healthcare records worldwide, supporting the use of standardized and uniform information (1-3). Nurses use standardized terminologies that are essential for identifying, processing and transmitting nursing practice data, making them visualizable and quantifiable (4, 5). The use of standardized nursing data improves the quality of nursing care and the accuracy of nursing records (6-8). Furthermore their use has a strong impact on public health in several countries (9) as it increases the safety and continuity of patient care, improves patient satisfaction and communication in the healthcare team and reduces healthcare costs as shown by different international studies (10-12).

At the same time, narrative and non-standardized nursing data are valuable and potentially meaningful resources about nursing care and are still the most represented data and clinical information in electronic health records (EHR) in many countries (13-16). This clinical information is often recorded in unstructured free-text and converting it to a structured format can be a time-consuming task that may not successfully capture all facets of the information (17). The important gain from the creation of structured data is the ability to manage and to mine clinical data in large volumes or across large time scales (18). Moreover, integration of diverse domain-associated datasets becomes critical to health care because semantic heterogeneity (SH) (i.e., the difference in meaning and interpretation of data elements) is detrimental to data interoperability (19), which is the extent to which systems can exchange and interpret shared data and be able to use it (20).

The interoperability of health information systems is crucial because it can improve the productivity and efficiency of healthcare to better serve public health worldwide as highlighted in a systematic review (21); furthermore, healthcare providers and researchers working with data must understand the importance of data mining (i.e., the set of techniques and methodologies that aim to extract useful information from large amounts of data) (22). Nurses play a pivotal role in this process as they are essential for data collection and generation of patient information (23).

Despite the growing need for standardized nursing data, most nursing records are represented without the use of a standard language, hindering an organized and systematic collection, analysis and interpretation of data that could be used for clinical practice, research and health policy development purposes (24, 23). Nursing records that are based on a combination of structured and unstructured data entry should ensure that the free-text data are available for reuse (25), whereas single free-text data, such as clinical narrative notes, would not allow comparison with other standardized languages. However, narrative notes present a challenge because they can take on many forms, as they reflect the nurse's perception of the patient's condition and can include a variety of highly telegraphic terms and many abbreviations (26). This variability poses some challenges by making it difficult to extract useful information from nursing notes due to their non-standardization; this current diversity in nursing vocabularies makes it impossible to compare nursing data over time, settings and populations (27, 28).

An automated text analysis can extract specific data and convert unstructured data into structured data from a corpus comprised of a significant amount of narrative data, and it can be run using natural language processing (NLP) and data mining techniques (29). NLP techniques as an artificial intelligence approach have been leveraged to extract information from clinical narratives in EHRs and to offer a strategy for integrating these approaches to provide structured reports for further computer processing (30).

According to a systematic review (31) which included and analyzed studies conducted in different continents, such as North America, Europe and Asia, NLP is currently the most widely used big data analytical technique in healthcare, and it is defined as a collection of syntactic and/or semantic rule or statistical-based processing algorithms that can be used to parse, segment, extract, or analyze text data (32). NLP algorithms can use defined language rules and relevant domain knowledge to perform syntactic processing (tokenization, sentence detection), extract information (e.g., convert unstructured text into a structured form), capture meaning (e.g., assign a concept to a word or group of words) and detect relationships (assign relationships between concepts) in natural language free-text (33). The implementation algorithm is vital for defining the role of NLP in data mining (34).

NLP techniques, by extracting documented nursing data in an unstructured form, can contribute to support clinical decisions, improving the research on patient outcomes and quality of care (29). International standardized nursing terminologies recognized by the American Nurses Association (ANA) (4), such as Clinical Care Classification System and International Classification for Nursing Practice, are widely used as structured data entry or reference terms with free-text nursing narratives assisted by NLP algorithms can be the source of data for comparing terms. This is important to determine their semantic equivalence through a strategy called cross-mapping, which allows the comparison of terms from different terminologies to determine their concept equivalence (17). Cross-mapping is the preferred nursing strategy to ensure the interoperability of healthcare data across terminologies in order to determine their semantic equivalence (28, 35). Of various cross-mapping solutions, the automation of term mapping can promote data integration, exchange and secondary use of clinical data (36). Using the cross-mapping method, research can bridge the gap between non-standardized nursing terms and standard terms (37) by measuring the contribution of nurses to patient outcomes, thus enabling a better understanding of clinical nursing findings and procedures. Using the cross-mapping method between non-standardized nursing terms and standard terms is an exciting possibility applicable in different health systems. This is why, many international studies have focused on text classification methods able to turn free-text nursing concepts into standard concepts (38) and, in particular, on mapping processes capable of checking whether a term of a terminology system matches or is comparable to a term in another terminology system (39-41). In these works, specific techniques were used to identify potentially synonymous nursing concepts expressed in free-text, turning them into standard concepts with one of three levels of cross-mapping (map, partial map and no map) (42), like other models that use text classification methods (e.g., Light Gradient Boosted model, [LightGBM] and Bidirectional Encoder Representations from Transformers model [BERT]) (43, 44). Unlike the models mentioned above, the String Metric-assisted Assessment of Semantic Heterogeneity (SMASH) was developed to support the evaluation of semantic heterogeneity among non-standardized clinical terms in other standardized languages. The SMASH is based on a string-assisted metric assessment and specifically used the *stringdist* command to compute pairwise string comparisons, similarity and distance, set to the Levenshtein distance (LD) and the Jaro-Winkler distance (JWD) (45). The SMASH method, used jointly with the NLP techniques, improves the data preparation to submit the unstructured free-text nursing activities in deep-learning architectures, such as deep neural networks (46, 47).

To our knowledge, no studies have been conducted on text analytics with NLP techniques and cross-mapping strategies with SMASH to process free-text nursing data recorded in EHRs, other than the models, found in the literature (48). The accuracy of text analytics with NLP techniques and the SMASH strategy could allow us to highlight and validate semantic differences between non-standardized nursing terms and other standardized languages, underlying the transformation process and the results connected to the transition from non-standardized nursing terms to standardized uniform nursing data. The aim of the study was 1) to identify unstructured free-text nursing activities recorded by nurses in EHRs with NLP techniques and 2) to map these nursing activities into standard nursing activities using the SMASH method.

Methods

Design

This is a retrospective study that utilized NLP techniques with a unidirectional mapping strategy.

Sample and Setting

A convenience sample (see also Data collection and variables) of nursing records from the Medicine, Neurology and Gastroenterology inpatient units of the Agostino Gemelli IRCCS University Hospital Foundation, Rome, Italy was considered for this study.

These inpatient units were selected due to the availability and accessibility of data by researchers; they were the first to be tested with the data analysis method proposed by researchers.

Records with a hospital length of stay (LOS) equal or above three days were included in the study since data about nursing activities were quantitatively richer.

Professional assessment instrument system

In the examined hospital, nursing records are documented using the professional assessment instrument (PAI) system. The PAI is the clinical nursing information system integrated in the hospital EHR. The PAI is used daily by nurses for documenting nursing care according to the steps of the nursing process (49, 50, 51).

Nurses can document their care using structured data such as nursing diagnoses, nursing activities, and non-standardized data such as nursing notes (52). PAI gives the opportunity to document every action made by the nurse both by selecting the corresponding standardized activity from a list of standardized and codified activities built-in in the PAI system and by typing additional free-text notes relating to each activity (non-standardized data that cannot be codified).

A total of 340 standard nursing activities is present on the PAI system. These activities are specific actions and behaviors adopted by nurses to enhance patient outcomes (e.g. enteral medication administration, assessment and monitoring of nutritional/hydration conditions, nasogastric tube management, vital signs measurement, urinary catheter management) (53).

Data collection and variables

Data about standardized nursing activities and non-standardized nursing activities were consecutively collected from July 2018 to December 2018. Non-standardized nursing activities (free-text from nursing notes) were extracted using NLP techniques (see next paragraphs) (54).

We included the following data fields found in the nursing documentation: short form, or the abbreviation (e.g., "RM"; "CPAP"; "PEV"; "RX TX"; "TC"; "ECG"; "EEG"; "ECO TSA"; "IC"; "PICC") and long form, or the spelled-out version of the abbreviation. Socio-demographic variables such as age and sex were also collected to describe the sample.

Data analysis

We applied the following three phases to analyze the data: 1) text summarization component with NLP techniques, 2) a consensus analysis by four experts to detect the category of word stems of unstructured free-text in nursing notes and 3) cross-mapping with SMASH.

***First phase**: Text summarization component with NLP techniques*

The aim of the first phase was to identify the most frequent word stems of unstructured free-text in nursing notes with NLP techniques. Unique words were extracted from unstructured free-text to allow an analysis of what nurses recorded. We implemented automatic text summarization in two steps (55, 56) (Figure1): 1. text pre-processing and 2. information extraction.

The text pre-processing included: a) syntactic analysis (i.e. unstructured free-text

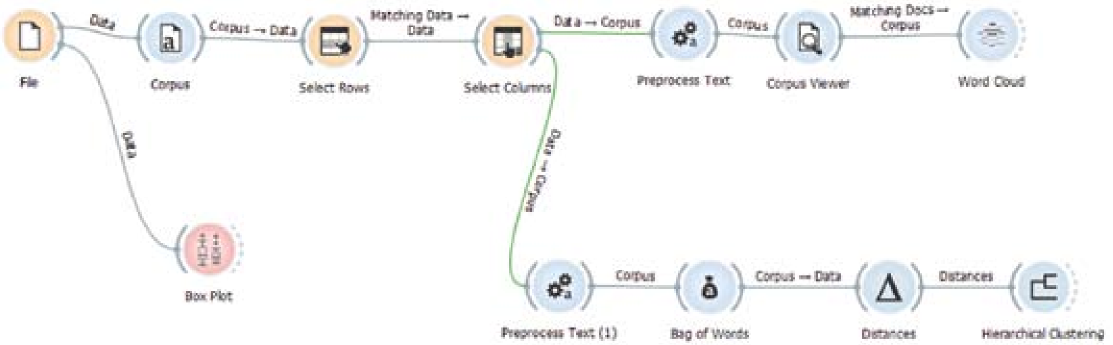


Figure 1 - ORANGE software v.3.22 and logical sequence with the programming language Python (Text pre-processing and Information extraction).

[a] *Note.* The NLP related widgets associated with the tokenization methodology application R@1-1 automatic evaluation metric.

in nursing notes were converted to word stems and were analyzed in corpus widget), b) tokenizer (i.e. unstructured free-text in nursing notes were analyzed as the sentence from the syntactic analysis into tokens with the methodology text mining application R@1-1, an automatic evaluation metric associated with pre-processing widget and free-text data were converted into arrays of sequences of numbers in regular expression - regexp 0.50-0.95 cut off and most frequent tokens > 100), c) semantic (i.e. unstructured free-text in nursing notes were divided into different classes, called part-of-speech (POS) tagging where a POS tag is a label assigned to each word in a text corpus of nursing notes), d) stop word removal (i.e. stop words and non-letter character punctuations were removed) and e) stemming (i.e. unstructured free-text in nursing notes were converted into arrays with frequency counts of each unique word stem and were analyzed in word cloud widget). The information extraction was the data cleaning process of the text summarization component, which analyzed the frequency of unstructured free-text in nursing notes, with a box plot function and hierarchical clustering widget.

All the analyses were performed with the ORANGE software v.3.22 (Copyright © University of Ljubljana), which is an open-source data visualization as a Python library, machine learning and data mining toolkit (57). The frequency counts of each unique word stem were calculated using a cut-off set in the tokenizer phase (see above). These word stems of unstructured free-text were considered in the second phase with consensus by four experts.

***Second phase:** A consensus analysis by four experts to detect the category of word stems of unstructured free- text nursing notes*

The aim of the second phase was to categorise the word stems of unstructured free-text nursing notes in macro areas (see below) to facilitate the cross-mapping method. To categorise the word stems, a consensus analysis was carried out by four experts (FD, AC, MV and MC) via the e-Delphi method in two rounds (58). Inclusion criteria for the panel of experts were as follows: high education in nursing (PhD or PhD student), more than five years of clinical experience, and expertise in nursing documentation.

Nursing notes contain words that make sense only in the context of a sentence, so the word stems identified were evaluated in the context of the sentence in which they were found. For example, the word stem ‘isolation’ was evaluated in the context of ‘to maintain contact isolation’.

In the first round, the word stems were analysed individually by each expert to understand whether the word stems referred to one of the following four macro areas: a) nursing diagnosis (e.g. patient problems), b) nursing activities (e.g. specific actions performed by nurses for nursing care), c) nursing handover (e.g. the shift report for patient care transferred from one nurse to another nurse) and d) patient-related nursing outcomes (e.g. the impact of nursing care on patient health). In the second round, the different categories identified by the four experts were compared and discussed in 25 one-hour online meetings and the level of agreement was classified as follows: ‘complete agreement’ (all experts agreed, 100% agreement); ‘partial agreement’ when three out of four experts reported the same analysis ($\geq$ 80% agreement); and ‘no agreement’ (consensus < 80%) when less than three experts agreed.

If ‘no agreement’ was found, a further round was carried out until a complete or partial agreement was reached. This phase was fundamental to consider only the word stems categorised in nursing activities—that is, unstructured free-text nursing activities—and to be submitted in the subsequent third phase: cross-mapping.

Third phase: *Cross-mapping unstructured free-text nursing activities into standard nursing activities with the SMASH method.*

The aim of the third phase was to map unstructured free-text nursing activities into standard nursing activities using the cross-mapping method with SMASH.

In this phase, one expert (MV) in cross-mapping method across terminologies analysed string comparison, similarity and distance (i.e. semantic heterogeneity) between the unstructured free-text nursing activities and PAI standard nursing activities with SMASH, which is a data-driven informatics method (45).

The SMASH method involved first calculating the distance using a *stringdist*, set to LD and JWD (59). JWD produces a score between 0–1 (1 = complete similarity and 0 = complete dissimilarity). The cut-offs identified to carry out the syntactic and semantic comparative analysis (with JWD) between unstructured free-text nursing activities and standard PAI nursing activities were as follows:

- map [0.80–1.00 JWD] with < 13 LD;
- partial map [0.50–0.79 JWD] >50% were repeated matches with two or more words being semantically-equivalent/syntactically-different, with < 13 LD;
- no map [0.0–0.49 JWD] with > 13 LD.

To analyse the cut-offs of cross-mapping and the application of the algorithm, unstructured free-text nursing activities were classified into true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) between *code_map* and *code_no_map*. Then, an F-measure was defined based on precision, recall and accuracy through which the reliability of cross-mapping was assessed. The F-measure was calculated using the harmonic mean of precision and recall to examine the performance of the automated SMASH procedure in generating accurate cross-mapping (F-measure = [2 x Recall x Precision]/[Recall + Precision]) and the area under the receiver operating characteristic (ROC) curve (60). The ROC was used to perform a binary classification of unstructured free-text nursing concepts between *code_map* and *code_no_map*. The area under the ROC curve was examined by plotting the TP rate (called *sensitivity*) against the FP rate (called 1 – *specificity*). In other words, the ROC curve studied the relationships between true alarms (hit rate) and false alarms. IBM ® SPSS statistical software (version 21) was used for data analyses.

Ethical aspects

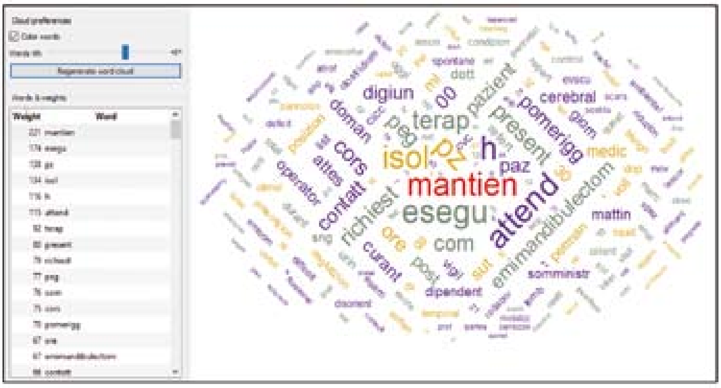
All data by which patients could be identified were anonymized and a unique numeric code was assigned to each record to be included in the data set. The protection and confidentiality of the data were guaranteed according to applicable privacy laws (61). The study was approved by the University Hospital Ethic Committee (research protocol N.0010375/20).

Results

According to the inclusion criteria a total of 491 patient records were assessed. The median hospital length of stay was 11.0 days (IQR = 11.0). The median age of the sample was 63 years (IQR = 24.0) and most of the patients were male 57.2%.

First phase: *Text summarization component with NLP techniques*

A total of 8491 tokens were extracted from free-text nursing notes and analysed with the text mining application R@1-1; subsequently, the identification of the tokens, 1,087 word stems, were calculated with the application of cut-off set in the tokenizer, semantic, stop word removal and stemming phases. These word stems were analyzed in word cloud widget (Figure 2).


Cloud preferences

☑ Color words

Words tilt

Regenerate word cloud

Words & weights

Weight	Word
221	mantien
174	esegu
128	pz
124	isol
116	h
115	attend
92	terap
80	present
79	richiest
77	peg
76	com
75	cors
75	pomerigg
67	ore
67	emimandibulectom
66	contatt

Figure 2 - Examples of word stems analyzed with word cloud widget of ORANGE.
[b] *Note*. The word stems were calculated with the application of cut-off set in the tokenizer, semantic, stop word removal and stemming phases.

Second phase: *A consensus analysis by four experts to detect the category of word stems of unstructured free- text nursing notes*

In this phase, of 1,087 word stems extracted, only those belonging to the macro area "nursing activities" were considered (see methods). A total of 548 word stems were categorized in nursing activities by four experts with 100% agreement. The e-Delphi results and the five most prevalent non-standardized nursing activities recorded by nurses are reported in Table 1 and Table 2, respectively.

Table 1 – The e-Delphi results of the final round by four experts

Macro Areas	Agreement Index %
Nursing Diagnosis	0.93
Nursing Activities	1.00
Nursing Handover	0.80
Patient-related Nursing Outcomes	0.86

[c] *Note.* Macro Areas had an agreement index ≥0.80.

Table 2 – The five most prevalent non-standardized nursing activities.

Word Stems	Sentence context: n (%)
Isolation	To maintain isolation; to remain isolation; contact isolation; pseudomonas isolation; preventive isolation: 392 (71.5%)
Liquid	To manage enteral liquids/fluids; to infuse liquid/physiological solution; to suspend liquids; intravenous liquids/fluids; to maintain liquids: 57 (10.4%)
Therapy	To administer therapy; to prescribe therapy; antibiotic therapy; to execute therapy; to practice therapy: 40 (7.3%)
CICC	To manage central catheter; to insert central catheter; to place central catheter; to maintain central catheter; to evaluate central catheter: 35 (6.4%)
SNG	To manage nasogastric tube; to replace nasogastric tube; to place nasogastric tube; to close nasogastric tube; to insert nasogastric tube: 24 (4.4%)

[d] *Note.* Word stems in the context of sentence analyzed.

Table 3 – Extract of cross-mapping between unstructured free-text nursing activities and PAI nursing standard activities.

Fully mapped activities [cut-off = 0.80-1.00]			
Unstructured free-text nursing activities	PAI Standard Nursing Activities (code)	JWD	LD
Bed bath completed	Bed bath completed (A.19.07)	0.99	1
Vital signs measured	Vital signs measured (A.02.01)	0.99	1
Bladder catheter management	Bladder catheter management (e.g., changing collection bag, patency check, skin care/check) (A.13.07)	0.99	1
Oxygen therapy with nasal cannula	Oxygen therapy with nasal cannula (A.11.13)	0.99	1
Stool sample taken	Sterile collection of a stool sample (A.09.13)	0.99	1
Nasogastric tube management	Nasogastric (SNG) or duodenal/ jejunal tube management (including proper insertion and care of the skin around the tube) (A.12.11)	0.99	1
Venous blood sampling for culture examination	Venous blood sampling for culture examination	0.99	1
CICC management	Central venous catheter management (including dressing and infusion set change) (A.05.06)	0.97	2
Electrocardiogram (ECG) performed and sent	Performing Electrocardiogram (ECG) (A.04.01)	0.97	2
Parenteral nutrition management	Parenteral preparation and administration of nutrients (parenteral nutrition) (A.12.14)	0.97	2
Irrigation performed	Instillations and/or irrigations of cavities, fistulas and ostomies (A.06.08)	0.95	3
Intravenous liquids	Intravenous perfusion management (e.g., flow rate monitoring, possible allergic reactions, site of needle skin emergence) (A.07.05)	0.95	3
Bladder washes	Bladder or intraurethral instillations or irrigations (A.13.14)	0.95	3
All hygiene care performed in bed	Hygiene care of one part of the body (A.19.05)	0.89	5
Urine sample taken for urine examination	Non-sterile collection of a urine sample (A.09.10)	0.89	5
American rectal probe replaced	Rectal probe introduction (A.14.07)	0.89	5
Enteral administration through PEG	Management of percutaneous endoscopic gastrostomy (PEG) (e.g., skin care around the tube, patency monitoring, irrigation) (A.12.12)	0.89	5
Administration foods in PEG	Preparation and administration of special foods by gastric tube or enteral pump (A.12.13)	0.89	5
Nasogastric tube insertion	Placement of nasogastric or duodenal tube (A.12.08)	0.84	6
AVP removed	Removal of cannula needle or butterfly needle (including dressing application) (A.05.05)	0.84	6
Therapy administered	Enteral administration of prescribed medications (A.07.01)	0.84	6
Partial mapped activities [cut-off = 0.50-0.79]			
Maintains night-time saturation	Oxygen saturation (SpO2) monitoring using a saturimeter (A.11.03)	0.79	7

Position changed several times	Decubitus changed every 1-2 hours (A.15.08)	0.74	8
Blood draws performed	Venous blood sampling (A.09.01)	0.74	8
Wound dressing performed	Wound/skin injury care (e.g., monitoring wound characteristics, dressing changes, wound cleansing, variation of the person's decubitus, using devices) (A.19.18)	0.74	8
Isolation procedure	Care and supervision of the person placed in a condition of protective isolation (A.20.13)	0.74	8
Patient's identification procedure	Application of an identification bracelet to the patient (A.20.06)	0.74	8
Intravenous drug administration	Parenteral administration of prescribed medications (A.07.02)	0.74	8
PICC management	Insertion of a peripherally inserted central catheter (PICC) (A.05.03)	0.69	9
CPAP management	Managing the mechanical ventilator and monitoring the ventilator-assisted fit (e.g., changing filters, system tubes) (A.11.22)	0.69	9
Urine Culture Test	Sterile collection of a urine sample (A.09.08)	0.59	11
Means of restraint applied	Physical restraint intervention according to medical prescription (e.g., physical restraint, area restriction) (A.20.10)	0.54	12
AVP positioned	Cannulation of a superficial vein with a cannula or butterfly needle (A.05.01)	0.54	12
Maintains conveen	External catheter (condom) management (e.g., change of collection bag, skin hygiene) (A.13.10)	0.54	12
Diaper and bed sheets changed	Assistance with incontinence or diarrhea (e.g., perineal hygiene, changing bedding/clothing, putting on diapers) (A.14.04)	0.54	12
No mapped activities **[cut-off = 0.00-0.49]**			
Diaper changed	Management of the excretory functions of the person through intimate aids with absorbent properties (diapers, absorbent garments) (A.13.11)	0.30	16
Suctioned when necessary	Maintenance of upper airway patency, suctioning of patient secretions (A.11.09)	0.30	16
Sputum sampling performed	/*	/	/
XDR (Extensively drug-resistant) performed	/*	/	/
Maintains graduated compression stockings	Nonmedicinal prevention of thrombosis (e.g., raising the affected limb above the level of the heart, wearing elastic stockings, promoting joint excursion movements) (A.15.12)	0.20	18

JWD - Jaro-Winker distance; LD - Levenshtein Distance.

[e] *Note.* Cut-offs identified to carry out the syntactic and semantic comparative analysis of unstructured free-text nursing activities with standard PAI nursing activities.
* no related PAI standard nursing activity.

Third phase: *Cross-mapping unstructured free-text nursing activities into standard nursing activities with the SMASH method.*

In this phase, 548 unstructured free-text nursing activities were mapped to PAI standard nursing activities and lexical similarities (with JWD) were the metrics used for the SMASH (Table 3). Lexical similarities between unstructured free-text nursing activities and PAI standard nursing activities were found with concepts identified for mapping. Of 548 concepts searched for automated lexical mapping (with JWD), 451 (82.3%) unstructured free-text nursing activities were mapped to PAI standard nursing activities (with <13 LD), 47 (8.7%) (of which 22 False negative and 25 False positive) were partial mapped (with <13 LD), while 50 (9.0%) were not mapped (with >13 LD). This automated mapping yielded a recall of 0.95%, a precision of 0.94%, an accuracy of 0.91% and an F-measure of 0.96. The F-measure (ranging from 0 to 1) indicates good reliability of this cross-mapping automated procedure. The area under the ROC curve of 0.97 (95% CI 0.96-0.98) shows that the automated procedure is accurate and performs to the cross-mapping method (reference line shown in Figure 3).

Discussion

To our knowledge, this is the first study aiming to cross-map unstructured free-text nursing activities into standard nursing activities using both NLP techniques and cross-mapping with the SMASH method. This study outlined the process of data analysis in three phases. The assessment of the semantic heterogeneity of non-standardized nursing terms was of paramount importance during the steps of this work.

In the first phase, the most frequent word stems of unstructured free-text nursing notes were analysed with the application of the text summarisation component with NLP techniques. In this phase, 1,087 word stems were calculated and this is an interesting result because the recognition of the word sense disambiguation of non-standardized nursing terms, abbreviations and acronyms

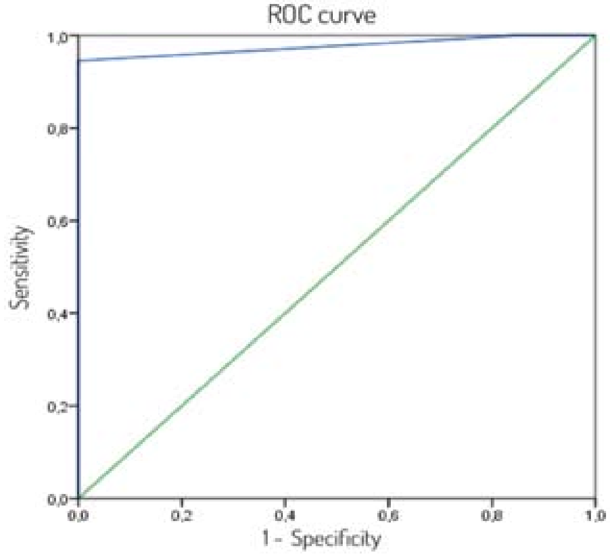


Figure 3 - The area under the ROC curve.
[f] *Note.* The cross-mapping method and the calculation of F-measure.

was crucial to prevent misinterpretation in the text pre-processing step of NLP (62, 63).

In the second phase, the word stems were categorised with a consensus analysis by four experts in macro areas to facilitate the cross-mapping method. The main result of the consensus analysis was an agreement index equal to 1.00. This score highlights complete agreement among all experts. Despite this result, there were some difficulties regarding the categorisation of some word stems. The evaluation of the word stems had to include the context of the sentence in which they were identified because a single concept can be expressed in many different words; likewise, it is possible to indicate with a single word different concepts in the nursing documentation (64, 65).

Free-text from nursing notes also included abbreviations and terms. The evaluation of the word stems highlights the results of the five most prevalent non-standardized nursing activities recorded by nurses: therapy administration, management of patient isolation for infection control, management of intravenous or enteral fluids, centrally inserted catheters and nasogastric tubes. The high prevalence of these activities in the hospital setting is supported by other studies (53, 66). Indeed, nurses in hospitals are mainly focused on therapy administration, monitoring patients' conditions and managing healthcare devices.

In the third phase, unstructured free-text nursing activities were cross-mapped into standard nursing activities using the SMASH method. Most of the free-text nursing activities were mapped into standard PAI nursing activities. Another Italian study found a similar result, cross-mapping free-text nursing activities recorded in paper-based documentation into the Nursing Intervention Classification (NIC) terminology (67).

A high prevalence of mapping between free-text nursing activities and standard nursing activities in the study hospital highlights that probably only a low percentage of nurses do not use the standard PAI nursing activities fully and tend to write in free-text instead of using the built-in standard nursing activities. Indeed, the total number of free-text nursing activities that we found in this study is much lower than the total number of structured PAI nursing activities that was found in another study with a similar sample size and period (53). However, we should consider that although EHRs are considered an opportunity to improve care, nurses often show dissatisfaction with their design and content (68, 52); this attitude could lead to their poor understanding and knowledge of the tools used. This study would therefore also place emphasis on the aspect of training in relation to EHR to make the documentation process more linear without repetition, allowing for better research work on these issues. The insertion of free-text in the EHR, when there are already standardized data with the same meaning, involves duplication of information and a waste of nursing time and resources that could be allocated to other activities; more in-depth knowledge of the EHRs used for nursing documentation, such as the PAI system, could also streamline the documentation process. In addition, we believe that the results obtained through this study can be used as a guide for assessing semantic heterogeneity among nursing data documented in the PAI system.

Finally, the cross-mapping method used in this study showed its accuracy. An F-measure of 0.96 and the area under the ROC curve of 0.97 remarked a weighted harmonic mean of recall and precision, thanks to the analysis of the lexical similarities between unstructured free-text nursing activities and PAI standard nursing activities. Concerning the high scores of recall, precision and accuracy indicated good reliability of this cross-mapping with the SMASH method. As this study can be considered a preliminary research for a possible future analysis, this framework was deemed suitable.

These results highlighted that the use of text analytics with NLP techniques, in the first phase, and cross-mapping strategies with SMASH, in the third phase, was crucial for processing free-text nursing data recorded in the EHRs, other than the models found in the literature (48). The study results, which were obtained using NLP techniques and the SMASH method, could help nurse researchers to better describe and understand all nursing activities for the medical areas and the mechanism of data interoperability.

The method allows the attribution of meaning to unstructured nursing data through a process of conversion and standardization within the medical record; such non-standardized narrative data becomes visible and potentially valuable for understanding the patient's complexity of care by giving meaning to nursing assessments and activities. In addition, by replicating the technique used in this study, nurse directors and educators could manage, monitor and teach the nursing documentation process in international institutional and academic settings.

Strengths and Study Limitations of the present experience

Strengths of the study were the use of an innovative technique (NLP with the SMASH method) to capture and standardise unstructured nursing generated data in EHRs (69) and the use of open-source tools for data science such as the ORANGE software. The main limitation of this study was the use of few inpatient units of a single hospital. This innovative technique could be replicated in studies including samples from different clinical settings; in fact, non-standard nursing activities are described by international literature as the predominant ones within EHRs (16). This strategy could be useful to confirm and generalise the results obtained from our study.

Conclusions

An automated method to support data analysis for comparing terms from different terminologies to later determine their concept equivalence through the SMASH method and NLP techniques is essential. It ensures the interoperability of healthcare data giving a greater understanding of the complexity of care related to hospitalized patients (70, 71). The encouraging results obtained through this study emphasize the possibility of replicating the method used in different care settings, such as paediatric, surgical and intensive care units. This aspect is fundamental to carry out also a semantic analysis during the translation process of free-text nursing data recorded in the EHRs.

Our proposed method was crucial for processing free-text nursing data recorded in the PAI system and for identifying various descriptions of nursing activities in nursing notes. We believe that it can be used to efficiently assist nurses in selecting the most appropriate nursing activities when they document the nursing process and to understand how to map and transform these unstructured free-text activities into standard activities, improving the conceptual clarity of nursing concepts (72). A practical impact of such a system would be that nurses save time and effort on tasks related to care documentation, which would result in more time to concentrate on the patient and deliver better care. Another outcome that could be improved is consistency and correctness in the use of concepts by nurses (73, 74). This research project highlights: the importance of an accurate extraction of nursing concepts (75, 76) and the potential ability of NLP to facilitate the advancement of the use of cross-mapping between the semantically-equivalent/syntactically-different terms with standard nursing activities to process or analyse information from unstructured free-text nursing activities (77).

Future research should focus on the implementation of this proposed method in different healthcare settings and countries to examine the congruence between non-standardized nursing terms and standardized nursing terminologies. It would allow the cross-mapping of unstructured free-text nursing activities with the use of a standardized nursing terminology, gaining a greater understanding of nursing care.

Funding: This work was funded by the Italian Center of Excellence for Nursing Scholarship, Rome, Italy
Competing interests: Not applicable
Conflicts of interest: none
Ethical Approval: Authorization obtained by the Catholic University Ethics Committee.
Acknowledgements: The authors would like to thank and to acknowledge the support received by the Agostino Gemelli IRCCS University Hospital Foundation, Rome, Italy for authorizing the study; the researchers involved for their participation and great support in data collection.

Riassunto

Linguaggio naturale e metodo String Metric-assisted Assessment of Semantic Heterogeneity per standardizzare le attività infermieristiche documentate in testo libero in ambito ospedaliero: uno studio retrospettivo

Introduzione. Gli infermieri generano dati assistenziali nelle cartelle cliniche elettroniche utilizzando diverse terminologie e sistemi di codifica. Lo scopo di questo studio è quello di identificare le attività infermieristiche registrate in testo libero dagli infermieri con tecniche di natural language processing (NLP) mappandole in attività infermieristiche standard utilizzando il metodo SMASH.

Disegno dello studio. Uno studio retrospettivo che ha utilizzato le tecniche di NLP con una strategia di mappatura unidirezionale, chiamata SMASH.

Metodi. Le attività infermieristiche in testo libero registrate nelle unità di degenza di Medicina, Neurologia e Gastroenterologia del Policlinico Gemelli, Roma, Italia sono state raccolte per 6 mesi nel 2018. I dati sono stati analizzati in tre fasi: a) la componente di riepilogo del testo con tecniche di NLP, b) una consensus analisi con quattro esperti per rilevare la categoria di appartenenza delle parole radici delle attività infermieristiche in testo libero e c) il cross-mapping con SMASH. Il metodo SMASH ha calcolato il confronto tra le stringhe, la somiglianza e la distanza delle parole tramite la distanza di Levenshtein (LD), di Jaro-Winker e i seguenti cut-off del cross-mapping: map completo [0.80-1.00] con < 13 LD, map parziale [0.50-0.79] con <13 LD e no map [0.0-0.49] con >13 LD.

Risultati. Durante il periodo di studio, sono state valutate 491 cartelle cliniche. Sono state rilevate 548 attività infermieristiche registrate in testo libero, di cui 451 attività (82.3%) sono state mappate in attività infermieristiche standard, 47 (8.7%) attività sono state parzialmente mappate, mentre 50 (9.0%) non sono state mappate. Questa mappatura automatizzata ha prodotto una sensibilità dello 0.95%, una precisione dello 0.94%, un'accuratezza dello 0.91%, e una misura F di 0.96. La misura F indica una buona affidabilità di questa procedura automatizzata nel cross-mapping.

Conclusioni. Sono state trovate somiglianze lessicali tra le attività infermieristiche in testo libero e le attività infermieristiche standard, il NLP con il metodo SMASH è un approccio possibile per estrarre dati relativi alle note infermieristiche scritte in testo libero.

References

1. Maas ML, Delaney C. Nursing process outcome linkage research: issues, current status, and health policy implications. Med. 2004; **42**(2 Suppl): 40-8. doi: 10.1097/01.mlr.0000109291.44014.cb.
2. D'Agostino F, Sanson G, Cocchieri A, et al. Prevalence of nursing diagnoses as a measure of nursing complexity in a hospital setting. J Adv Nurs. 2017; **73**(9): 2129-42. doi: 10.1111/jan.13285.
3. Galatzan BJ, Carrington JM. Examining the meaning of the language used to communicate the nursing hand-off. Res Nurs Health. 2021; **44**(5): 833-43. doi: 10.1002/nur.22175.
4. Tastan S, Linch GCF, Keenan GM, et al. Evidence for the existing American Nurses Association-recognized standardized nursing terminologies: a systematic review. Int J Nurs Stud. 2014; **51**(8): 1160-70. doi: 10.1016/j.ijnurstu.2013.12.004.
5. Häyrinen K, Saranto K. The use of nursing terminology in electronic documentation. Stud Health Technol Inform. 2009; **146**: 342-6. doi: 10.3233/978-1-60750-024-7-342.
6. D'Agostino F, Zeffiro V, Vellone E, et al. Cross-Mapping of Nursing Care Terms Recorded in

Italian Hospitals into the Standardized NNN Terminology. Int J Nurs Knowl. 2020; **31**(1): 4-13. doi: 10.1111/2047-3095.12200.

7. De Groot K, De Veer AJE, Paans W, Francke AL. Use of electronic health records and standardized terminologies: A nationwide survey of nursing staff experiences. Int J Nurs Stud. 2020; **104**: 103523. doi: 10.1016/j.ijnurstu.2020.103523.
8. Sanson G, Vellone E, Kangasniemi M, Alvaro R, D'Agostino F. Impact of nursing diagnoses on patient and organisational outcomes: a systematic literature review. J Clin Nurs. 2017; **26**(23-24): 3764-3783. doi: 10.1111/jocn.13717.
9. Rabelo-Silva ER, Dantas Cavalcanti AC, Ramos Goulart Caldas MC, Lucena AF, Almeida MA, Linch GF, da Silva MB, Müller-Staub M. Advanced Nursing Process quality: Comparing the International Classification for Nursing Practice (ICNP) with the NANDA-International (NANDA-I) and Nursing Interventions Classification (NIC). J Clin Nurs. 2017; **26**(3-4): 379-387. doi: 10.1111/jocn.13387.
10. Ali S, Sieloff CL. Nurse's use of power to standardise nursing terminology in electronic health records. J Nurs Manag. 2017; **25**(5): 346-353. doi: 10.1111/jonm.12471.
11. Chae S, Oh H, Moorhead S. Effectiveness of Nursing Interventions using Standardized Nursing Terminologies: An Integrative Review. West J Nurs Res. 2020; **42**(11): 963-973. doi: 10.1177/0193945919900488.
12. Saba VK, Arnold JM. Clinical care costing method for the Clinical Care Classification System. Int J Nurs Terminol Classif. 2004; **15**(3): 69-77. doi: 10.1111/j.1744-618x.2004.tb00002.x.
13. Tubaishat A. The effect of electronic health records on patient safety: A qualitative exploratory study. Inform Health Soc Care. 2019; **44**(1): 79-91. doi: 10.1080/17538157.2017.1398753.
14. Tsai CH, Eghdam A, Davoody N, et al. Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content. Life. 2020; **10**(12): 327. doi: 10.3390/life10120327.
15. Yang X, Bian J, Fang R, et al. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc. 2020; **27**(1): 65-72. doi: 10.1093/jamia/ocz144.
16. Bowles KH, Potashnik S, Ratcliffe SJ, et al. Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators. J Nurs Adm. 2013; **43**(6): 355-60. doi: 10.1097/NNA.0b013e3182942c3c.
17. Szostak J, Ansari S, Madan S, et al. Construction of biological networks from unstructured information based on a semiautomated curation workflow. Database (Oxford) 2015; 1-14. doi: 10.1093/database/bav057.
18. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017; **73**: 14-29. doi: 10.1016/j.jbi.2017.07.012.
19. Livingston KM, Bada M, Baumgartner WA, Hunter LE. KaBOB: ontology-based semantic integration of biomedical databases. BMC Bioinformatics. 2015; **16**: 126. doi: 10.1186/s12859-015-0559-3.
20. Otokiti A. Using informatics to improve healthcare quality. Int J Health Care Qual Assur. 2019; **32**(2): 425-430. doi: 10.1108/IJHCQA-03-2018-0062.
21. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst. 2018; **42**(11): 214. doi: 10.1007/s10916-018-1075-6.
22. Pine KH. The qualculative dimension of healthcare data interoperability. Health Informatics J. 2019; **25**(3): 536-548. doi: 10.1177/1460458219833095.
23. Urquhart C, Currell R, Grant MJ, Hardiker NR. Nursing record systems: effects on nursing practice and healthcare outcomes. Cochrane Database Syst Rev. 2009; **1**: CD002099. doi: 10.1002/14651858.CD002099.pub2. Update in: Cochrane Database Syst Rev. 2018; **5**: CD002099.
24. D'Agostino F, Zega M, Rocco G, Luzzi L, Vellone E, Alvaro R. Impact of a nursing information system in clinical practice: a longitudinal study project. Ann Ig. 2013; **25**(4): 329-41. doi: 10.7416/ai.2013.1935.
25. Wulff A, Mast M, Hassler M, Montag S, Marschollek M. Designing an openEHR-Based Pipeline for Extracting and Standardizing Unstructured Clinical Data Using Natural Language Processing. Methods Inf Med. 2020; **59**(2): 64-78. doi: 10.1055/s-0040-1716403.
26. Long WJ. Parsing Free Text Nursing Notes. AMIA Annu Symp Proc. 2003; 917.

27. Elfrink V, Bakken S, Coenen A, McNeil B, Bickford C. Standardized nursing vocabularies: a foundation for quality care. Semin Oncol Nurs. 2001; **17**(1): 18-23. doi: 10.1053/sonu.2001.20415.
28. Goossen W. Cross-mapping between three terminologies with the international standard nursing reference terminology model. Int J Nurs Terminol Classif. 2006; **17**(4): 153-64. doi: 10.1111/j.1744-618X.2006.00034.x.
29. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs. 2009; **27**(4): 215-23. doi: 10.1097/NCN.0b013e3181a91b58.
30. Hyun S, Bakken S, Friedman C, Johnson SB. Natural language processing challenges in HIV/AIDS clinic notes. AMIA Annu Symp Proc. 2003; 872.
31. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inform. 2018; **114**: 57-65. doi: 10.1016/j.ijmedinf.2018.03.013.
32. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform. 2019; **125**: 37-46. doi: 10.1016/j.ijmedinf.2019.02.008.
33. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. Methods. 2015; **74**: 97-106. doi: 10.1016/j.ymeth.2015.01.015.
34. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020; **145**(2): 463-9. doi: 10.1016/j.jaci.2019.12.897.
35. Torres FBG, Gomes DC, Hino AAF, Moro C, Cubas MR. Comparison of the Results of Manual and Automated Processes of Cross-Mapping Between Nursing Terms: Quantitative Study. JMIR Nurs. 2020; 9; **3**(1):e18501. doi: 10.2196/18501.
36. Lu F, Park HT, Ucharattana P, Konicek D, Delaney C. Nursing outcomes classification in the systematized nomenclature of medicine clinical terms: a cross-mapping validation. Comput Inform Nurs. 2007; **25**(3): 159-70. doi: 10.1097/01.NCN.0000270042.22164.21.
37. Sun JY, Sun Y. A system for automated lexical mapping. J Am Med Inform Assoc. 2006; **13**(3): 334-43. doi: 10.1197/jamia.M1823.
38. Forsvik H, Voipio V, Lamminen J, Doupi P, Hypponen H, Vuokko R. Literature review of patient record structures from the physician's perspective. J Med Syst. 2017; **41**(2): 29. doi: 10.1007/s10916-016-0677-0.
39. Kieft RAMM, Vreeke EM, de Groot EM, et al. Mapping the Dutch SNOMED CT subset to Omaha System, NANDA International and International Classification of Functioning, Disability and Health. Int J Med Inform. 2018; **111**: 77-82. doi: 10.1016/j.ijmedinf.2017.12.025.
40. Junglyun K, Yingwei Y, Tamara Goncalves Rezende M, Gail K. An examination of the coverage of the SNOMED CT coded nursing problem list subset. JAMIA Open. 2019; **2**(3): 386-91. doi: 10.1093/jamiaopen/ooz023.
41. Vis L, Koole S, Goossen A, Huisman H, Goossen W. Semantic Cross-Mapping Execution of Data in the Perinatal Registry of the Netherlands. Stud Health Technol Inform. 2020; **273**: 117-22. doi: 10.3233/SHTI200625.
42. Kim TY, Hardiker N, Coenen A. Inter-terminology mapping of nursing problems. J Biomed Inform. 2014; **49**: 213-20. doi: 10.1016/j.jbi.2014.03.001.
43. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017; 3147-55.
44. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL. 2019; **19**(1): 4171-86.
45. Brown W, Weng C, Vawdrey DK, Carballo-Diéguez A, Bakken S. SMASH: A Data-driven Informatics Method to Assist Experts in Characterizing Semantic Heterogeneity among Data Elements. AMIA Annu Symp Proc. 2017; **10**: 1717-26.
46. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018; **19**(6): 1236-46. doi: 10.1093/bib/bbx044.
47. Simões MF, Silva G, Pinto AC, Fonseca M, Silva NE, Pinto RMA, Simões S. Artificial neural networks applied to quality-by-design: From formulation development to clinical outcome. Eur J Pharm Biopharm. 2020; **152**: 282-95. doi: 10.1016/j.ejpb.2020.05.012.
48. Liu LG, Grossman RH, Mitchell EG, et al. A deep database of medical abbreviations and acronyms for natural language processing. Sci Data. 2021; **8**(1): 149. doi: 10.1038/s41597-021-00929-4.

49. Cocchieri A, Di Sarra L, D'Agostino F, et al. Sviluppo e implementazione di un sistema informativo infermieristico pediatrico in ambito ospedaliero: il PAI pediatrico [Development and implementation of pediatric and neonatal nursing information system in an hospital setting: the pediatric PAI]. Ig Sanita Pubbl. 2018; **74**(4): 315-28.
50. D'Agostino F, Zega M, Rocco G, et al. Impact of a nursing information system in clinical practice: a longitudinal study project. Ann Ig. 2013; **25**(4): 329-41. doi: 10.7416/ai.2013.1935.
51. Zega M, D'Agostino F, Bowles KH, et al. Development and validation of a computerized assessment form to support nursing diagnosis. Int J Nurs Knowl. 2014; **25**(1): 22-9. doi: 10.1111/2047-3095.12008.
52. D'Agostino F, Vellone E, Tontini F, Zega M, Alvaro R. Sviluppo di un sistema informativo utilizzando un linguaggio infermieristico standard per la realizzazione di un Nursing Minimum Data Set [Development of a computerized system using standard nursing language for creation of a nursing minimum data set]. Prof Inferm. 2012; **65**(2): 103-9.
53. Sanson G, Alvaro R, Cocchieri A, et al. Nursing Diagnoses, Interventions, and Activities as Described by a Nursing Minimum Data Set: A Prospective Study in an Oncology Hospital Setting. Cancer Nurs. 2019; **42**(2): 39-47. doi: 10.1097/NCC.0000000000000581.
54. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019; **26**(4): 364-79. doi: 10.1093/jamia/ocy173.
55. Bacchi S, Gluck S, Tan Y, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. Intern Emerg Med. 2019; **15**(6): 989-95. doi: 10.1007/s11739-019-02265-3.
56. Cook MJ, Yao L, Wang X. Facilitating Accurate Health Provider Directories Using Natural Language Processing. BMC Med Inform Decis Mak. 2019; **19**(3): 80. doi: 10.1186/s12911-019-0788-x.
57. Toplak M, Birarda G, Read S, et al. Infrared Orange: Connecting Hyperspectral Data with Machine Learning. Technical Reports. 2017; **30**(4): 40-5. Available on: https://en.wikipedia.org/wiki/Orange_(software) [Last accessed: 2021, May 05]. doi: 10.1080/08940886.2017.1338424.
58. Hong QN, Pluye P, Fàbregues S, et al. Improving the content validity of the mixed methods appraisal tool: a modified e-Delphi study. J Clin Epidemiol. 2019; **111**: 49-59.e1. doi: 10.1016/j.jclinepi.2019.03.008.
59. Available on: https://it.functions-online.com/levenshtein.html [Last accessed: 2021, June 04].
60. Kim TY. Automating lexical cross-mapping of ICNP to SNOMED CT. Inform Health Soc Care. 2016; **41**(1): 64-77. doi: 10.3109/17538157.2014.948173.
61. Regolamento generale per la protezione dei dati personali del 24 maggio 2016, n. 679. General Data Protection Regulation o GDPR, normativa europea in materia di protezione dei dati.
62. Bjarnadottir RI, Lucero RJ. What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A Text Mining Study. EGEMS. 2018; **6**(1): 1-8. doi: 10.5334/egems.237.
63. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. Int J Med Inform. 2019; **129**: 184-8. doi: 10.1016/j.ijmedinf.2019.06.008.
64. Le QV, Mikolov T. Distributed representations of sentences and documents. Int Conf Mach Learn. 2014; 1188-96.
65. Zeffiro V, Sanson G, Vanalli M, et al. Translation and cross-cultural adaptation of the Clinical Care Classification system. Int J Med Inform. 2021; **153**: 104534. doi: 10.1016/j.ijmedinf.2021.104534.
66. Kang MJ, Dykes PC, Korach TZ, et al. Identifying nurses' concern concepts about patient deterioration using a standard nursing terminology. Int J Med Inform. 2020; **133**: 104016. doi: 10.1016/j.ijmedinf.2019.104016.
67. Lavin MA, Harper E, Barr N. Health Information Technology, Patient Safety, and Professional Nursing Care Documentation in Acute Care Settings. Online J Issues Nurs. 2015; **20**(2): 6. doi: 10.3912/OJIN.Vol20No02PPT04.
68. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Adv Neural Inf Process Syst. 2015; 649-57.
69. Bravetti C, Cocchieri A, D'Agostino F, Alvaro R, Zega M. The assessment of the complexity of care through the clinical nursing information system in clinical practice: a study protocol. Ann Ig. 2017; **29**(4): 273-80. doi: 10.7416/ai.2017.2155.

70. Bravetti C, Cocchieri A, D'Agostino F, et al. A nursing clinical information system for the assessment of the complexity of care. Ann Ig. 2018; **30**(5): 410-20. doi: 10.7416/ai.2018.2241.
71. Moen H, Hakala K, Peltonen LM, et al. Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. J Am Med Inform Assoc. 2020; **27**(1): 81-88. doi: 10.1093/jamia/ocz150.
72. Heidarizadeh K, Rassouli M, Manoochehri H, Tafreshi MZ, Ghorbanpour RK. Effect of electronic report writing on the quality of nursing report recording. Electron Physician. 2017; **9**(10): 5439-45. doi: 10.19082/5439.
73. Häyrinen K, Lammintakanen J, Saranto K. Evaluation of electronic nursing documentation-nursing process model and standardized terminologies as keys to visible and transparent nursing. Int JMed Inform. 2010; **79**(8): 554-64. doi: 10.1016/j.ijmedinf.2010.05.002.
74. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif Intell Med. 2015; **65**(2): 155-66. doi: 10.1016/j.artmed.2015.04.007.
75. Mitchell B, Petrovskaya O, McIntyre M, Frisch N. Where is nursing in the electronic health care record? Stud Health Technol Inform. 2009; **143**: 202-06. doi: 10.3233/978-1-58603-979-0-202.
76. Westra BL, Latimer GE, Matney SA, et al. A national action plan for sharable and comparable nursing data to support practice and translational research for transforming health care. J Am Med Inform Assoc. 2015; **22**(3): 600-07. doi: 10.1093/jamia/ocu011.
77. Johnson SG, Pruinelli L, Westra BL. Machine Learned Mapping of Local EHR Flowsheet Data to Standard Information Models using Topic Model Filtering. AMIA Annu Symp Proc. 2020; **4**: 504-13.

Corresponding author: Mariangela Vanalli, RN, MsN, PhD, Department of Biomedicine and Prevention, University of Rome Tor Vergata, Via Montpellier, 1 00133 Rome, Italy
e-mail: mariangelavanalli@gmail.com
ORCID: 0000-0002-47461264