Data Quality and Completeness in a Web Stroke Registry as the Basis for Data and Process Mining

Giordano Lanzola, PhD[1]; Enea Parimbelli, MS[1]; Giuseppe Micieli, MD[2]; Anna Cavallini, MD[2] and Silvana Quaglini, PhD[1]*

[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

[2]IRCCS Foundation C. Mondino, Pavia, Italy

Submitted September 2013. Accepted for publication February 2014.

ABSTRACT

Electronic health records often show missing values and errors jeopardizing their effective exploitation. We illustrate the re-engineering process needed to improve the data quality of a web-based, multicentric stroke registry by proposing a knowledge-based data entry support able to help users to homogeneously interpret data items, and to prevent and detect treacherous errors. The re-engineering also improves stroke units coordination and networking, through ancillary tools for monitoring patient enrollments, calculating stroke care indicators, analyzing compliance with clinical practice guidelines, and entering stroke units profiles. Finally we report on some statistics, such as calculation of indicators for assessing the quality of stroke care, data mining for knowledge discovery, and process mining for comparing different processes of care delivery. The most important results of the re-engineering are an improved user experience with data entry, and a definitely better data quality that guarantees the reliability of data analyses.

Keywords: data acquisition, human computer interaction, disease registry, stroke unit, statistical indicators

1. INTRODUCTION AND BACKGROUND

The analysis of clinical data collected over a large population sample affected by the same disease has a potential impact on streamlining the management of that disease and on improving its outcome [1]. Specialized databases are therefore being designed, usually referred to as Disease Registries (DRs), that are becoming very popular as a means of conducting investigations over diseases [2-3]. Arts et al. [4] have informally categorized DRs as "a systematic collection of a clearly defined set of health and demographic data for patients with specific health characteristics, held in a central database for a predefined

* Corresponding author: Silvana Quaglini, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100 Pavia, Italy. Phone: 0039-0382-985981. Fax: 0039-0382-985373. Email: silvana.quaglini@unipv.it.Other authors':giordano. lanzola@unipv.it;enea.parimbelli@gmail.com;giuseppe.micieli@mondino.it;anna.cavallini@mondino.it

purpose". Among those purposes, there are activities such as epidemiological research, care process monitoring, data mining, and process mining, which require large amounts of data. Those activities may highly benefit from DRs since they are conceived to collect large sets of data in a relatively short time. However, despite the big potential offered by DRs, a major bottleneck for their effective exploitation lies in the quality of collected data. Sometimes, in fact, this is insufficient to yield meaningful results from statistical analysis. That is why data acquisition represents a key issue in the overall process and currently shows several weaknesses. First, due to many logistic and organizational reasons, primary data collection for DRs is still largely accomplished using paper and put into electronic format at a later time. This happens since computer applications at hospitals are highly fragmented and unable to interoperate or exchange data among each other. Even centers equipped with Electronic Medical Records (EMRs) cannot automatically feed DRs, and the data collection process requires instead a double data-entry[1] step. This is time-consuming and is also the cause of many errors, but we still have to cope with this problem on the mid-term [6-7]. Second, even though a suitable set of validations could be accomplished upon electronic data entry to enforce the required quality, this check is often lacking or very poor, causing the well-known "garbage-in garbage-out" problem [8] which renders the whole study unreliable. "Pitfalls of going electronic" are well analyzed by Hartzband [9], who discusses the pros and cons of EMRs versus paper-based clinical charts, and the importance of accounting for human factors in the design and development of human-computer interfaces for clinical applications [10].

This paper deals with data quality of a stroke registry. Stroke is among the leading causes of death and chronic impairments, and it also accounts for major expenditures in the health care budgets of the western countries [11]. Stroke Unit (SU) care represents a major achievement in the management of patients affected by ischemic stroke with respect to their treatment in general wards [12-13]. Despite this evidence, SUs were not common until a few years ago, and still now they are not homogenously spread in Italy [14-15]. From the year 2000, in order to switch to a better organized inpatient care, and improve the compliance with clinical practice guidelines, the Italian National Health Service started endowing SUs in major hospitals in Lombardia, a region located in northern Italy. However, it became clear almost immediately that some efforts were needed for harmonizing the delivery of health care processes across the whole region. To this aim, the Stroke Unit Network (SUN) Lombardia was launched as a network of excellence among 40 participating sites in order to identify patients' needs and improve the exchange of clinical and therapeutic information among the health care operators involved [16].

1.1. The First Version of the SUN Registry and its Problems

As a foundation for any activity to be carried out by SUN Lombardia, a first version of the SUN Registry was developed as a web application in 2007 by an external consulting company. That registry was set up to collect data about patients diagnosed with acute ischemic stroke, Transient Ischemic Attacks (TIAs) and cerebral

[1] With the term "double data-entry" we do not refer to the practice of entering data twice in order to detect errors [5], but rather to the inability of automatically transferring data already existing in some other repository into a different database.

hemorrhages. It was aimed at coordinating and streamlining the actions of all the participating units fostering the adherence to stroke care guidelines by monitoring the quality, efficacy and efficiency of the stroke care in different centers [17]. Data were retrospectively collected from the clinical charts (either in electronic or paper format, depending on the center) selecting patients on the basis of their hospital discharge codes. The registry included patient-level information arranged in a demographic section and the four main events reflecting the entire stroke path, namely Emergency, Stroke Unit Admission, Discharge and Follow-up. The common stroke scales, namely the National Institutes of Health Stroke Scale (NIHSS), Barthel and Rankin, were used to assess the patient status at various phases, since they are important care process indicators.

By December 31st, 2008, over 6,000 patient cases were available for statistical analyses. Unfortunately, the problem of poor data quality was immediately spotted, since this first release of the SUN Registry was missing even the simplest data entry quality controls. The extensive use of free text data was a further limitation to the possible statistics. Thus, the accomplishment of the study has required a careful data cleaning and pre-processing, before the effectiveness of the registry could be shown [18]. Moreover, the website only provided the data entry Graphical User Interface (GUI) and a data export facility in terms of a flat spreadsheet format, so that statistics or reports were neither immediately available to the end users nor even to the registry coordinator. As a consequence, differences among centers, concerning both the data quality and the patients' enrollment rate, only emerged through *a-posteriori* analyses, making it difficult to promptly undertake any corrective action effectively. Poor data quality and large differences on the number of patients enrolled at the participating centers called not only for a tighter quality control, but also for a more functional website, with additional tools to make sure that all the users were properly filling out the registry.

Due to the problems experienced with its usage, as only partially introduced in the above discussion, in 2009, it was decided to completely re-engineer the SUN Registry, undertaking a major effort (it took six months) similar to the one accomplished on another stroke registry in Texas [19]. This paper illustrates our efforts in accomplishing such re-engineering, and the results achieved in terms of data quality and data analysis.

2. METHODS

In this section, we first review and analyze the main causes that contribute to spoiling data acquisition, since avoiding those was the primary goal of the re-engineering of the SUN Registry. Then we provide some suggestions on the design of a DR and EMRs in general, which could improve data acquisition.

2.1. Analysis of the Causes of Poor Data Quality

There are many reasons, as described below, that could impair the quality of data acquired electronically. Certainly, there is a trade-off between implementing a rigorous data control, also considering the future usage of the collected data, and providing an easy-to-use and fast GUI.

2.1.1. Poor Data Entry Control

GUIs implement quality controls for data entry at different levels. In the simplest one, there are:

- procedural checks that enforce data types, unit of measurements or consistency of the admissible value ranges;
- input masks guaranteeing that the values introduced comply with a predefined format (e.g., the social security number in the US or the taxpayer code in Italy);
- dropdown menus constraining the user input to a small set of admissible values (e.g., "Male" or "Female" for gender) disallowing free text.

On top of those syntactical rules, some systems implement semantic rules enforcing consistency among separate data chunks. An example is given by temporal validation constraints set up among multiple dates or by the deactivation of input items that are not compatible with the information already entered (e.g., the menopausal status for a male patient).

However, the most treacherous errors in data entry cannot be avoided with those methods, since their detection can occur only exploiting the domain medical knowledge which encapsulates the relationships among data. In fact, some correlated data may be entered at different times, or even be spread over different sections of the patient record, making it difficult for the user or for the data analyst to identify incoherencies. An example of those difficult-to-detect errors is given by a therapy which is not suitable for a given diagnosis (e.g., when a patient record simultaneously reports "anticoagulant drug" as the therapy and "cerebral hemorrhages" as the diagnosis). Of course, those data could be actually true, and the inconsistency could be due to a medical error (anticoagulants are detrimental for hemorrhagic patients). In any case, this is a situation requiring a double check by the system. Another example comes from the common use of "0" to represent a missing value for a numeric item. This causes an ambiguity when the data item includes "0" among its admissible values.

Other sources of errors during data-entry are related to organizational issues. For example, to overcome the physicians' resistance to electronic data entry [20], healthcare organizations may enroll dedicated people who are mostly non-medical for that task. Literature shows conflicting findings on this subject: on one hand, an inappropriate level of training of non-medical staff seems to be one of the major causes of data input errors and missing values [21,22]; on the other hand, there is some evidence that variability increases when data abstraction and data entry are performed by clinicians rather than by the clerical staff [23]. On this basis, we might argue that a non-clinical, but well trained, clerical staff is the best option.

Finally, real-time data entry is still mostly a technological problem, since mobile technology has not widely entered the clinical routine yet, and wireless networks and devices are more demanding in terms of patient safety. Thus, keyboard data entry is not always available at the patient bedside, and alternate modes, such as voice recognition, are not ripe enough for regular use in noisy environments such as hospital wards [24]. In summary, it is difficult to avoid errors and missing data because of lack of sophisticated controls, and also because a patient record is filled out by different people in challenging environments.

2.1.2. Lack of a Shared Cognitive Model of the Domain
Particularly in DRs that require input from several centers, data are entered by different people with a varied background, expertise and skills, and are gathered from clinical charts and additional documents whose format may vary across centers. As a consequence, the end users often misunderstand what information is actually required by a given input item. This happens because labels in entry forms are very concise and sometimes turn out to be even ambiguous for those end users that were not involved in the system development and therefore do not share the same cognitive model of the data. To overcome that problem, the system should make explicit and share among all users the domain specific knowledge adopted to model the data being acquired and their relationships. This knowledge could be represented in the system through a semantic network and its relationships should be used to automatically derive rules for enforcing consistency checks.

2.1.3. Gap Between Design and Exploitation of a System
It may happen that a system is designed and developed for a specific purpose, but after some time it starts to be used also for other ones. A gap between the original design and the actual exploitation may lead to data misuse and erroneous results. This happens since the purpose of a system determines which data are to be entered, their level of detail, the most appropriate temporal granularity and even the required quality, from which all the input controls will depend. For example, imagine a system aimed at simply transforming a paper-based clinical chart (usually relying on free text) into an electronic one, maintaining the same data structure. If then we ask for automatic data analysis, a statistician will have to face many problems [25,26].

2.1.4. Lack of Contextual Information
Healthcare records rarely include any contextual information, that is information about the environment where data are collected. However, such information could be useful for data interpretation. For example, data can be exploited to identify non-compliance with clinical practice guidelines [27-28]. In similar studies, it is very important to distinguish between non-compliances due to a physician's choice, and those due to lack of resources (specific skills or devices required to perform a recommended action). Thus, organizational data, with particular emphasis on those describing human or technological resources, should be saved along with clinical data. Timestamps are also crucial, since resources vary with time. For instance, a hospital could have a CT-scan available only during the day. If, after some time, an additional CT-scan and/or a night-time operator become available, an increasing trend of the CT-scans performed should not be surprising. Those considerations are also important when comparing processes and healthcare indicators among different hospitals.

2.1.5. Lack of Incentives for Users
Although rare, incentives could be extremely useful to lower the barriers against the use of computerized systems, and to foster the accomplishment of data collection in a timely and effective fashion. Users may be rewarded by facilitating their routine tasks,

e.g., automatically obtaining some statistics or reports, sharing the authorship of scientific papers, being proactive by providing feedback to the developers of the system in order to improve the system itself, and eventually feeling part of a *learning* community.

2.2. The Involvement of End Users in the Requirement Analysis

The first precondition to acquire good quality data is to pay a special attention to the GUI design. Unfortunately, physicians are not much interested in the design phase, and quite often, they let computer scientists and technicians decide for them. Being end users, physicians simply request easy and fast data entry procedures, and it is also not uncommon that they ask to replicate the same layout of their previous paper-based clinical charts. Nevertheless, a careful and shared preliminary analysis is mandatory to anticipate as much as possible any type of future exploitation of the acquired data. In fact, the impossibility to accomplish the necessary statistics is usually realized at a later time when it's too late for revising the model. In the following paragraphs, we describe some examples that we used to illustrate, to the SUN Registry coordinator and his staff, how different ways of acquiring the same data, in spite of looking similar at first sight, may instead deeply affect the information they convey.

We start with binary data, the acquisition of which is typically accomplished using single checkboxes, radio buttons or dropdown menus. Let us consider the stroke risk factor *Atrial Fibrillation*, and the five different possibilities of acquiring it, illustrated in Figure 1. Solution *(a)* is probably the simplest and the fastest one from the data entry operator's point of view, but it turns out to be ambiguous. In fact, if the box is not checked, we cannot distinguish between "No" (i.e., risk factor not present) or "Unknown" (i.e., no information available). Nevertheless it may be important to differentiate among those two cases in order to interpolate or filter out missing values before performing statistical analyses.

The solutions *(b)* and *(c)* are equivalent from the information point of view, and choosing between them is just a matter of layout. If the operator does not check anything, the information is assumed to be "Unknown". Please note however that

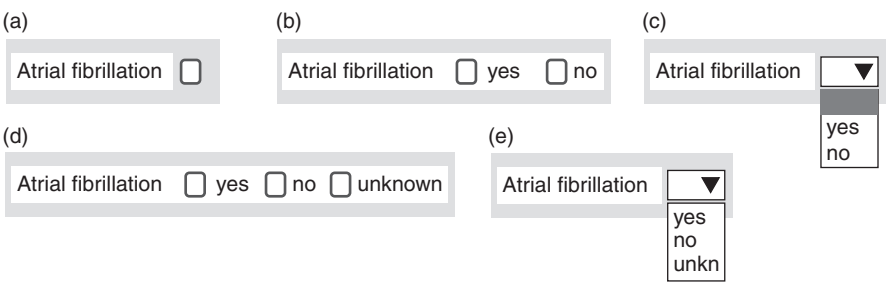


Figure 1. Five different ways of entering binary information, such as the presence of a risk factor.

solution *(c)* requires two clicks (i.e., a first one to reveal the dropdown menu and another one to select the actual item) instead of one. Despite the fact that *(b)* and *(c)* provide different states for all the three information values (i.e., "Yes", "No", "Unknown") they still conceal a tricky error. Since the "Unknown" option is represented by the implicit absence of a click over "Yes" or "No", if the operator skips that item either because of inattentiveness or interruption by any other intervening task, the form would be validated as well. Thus an "Unknown" value would be assumed even though a value does exist. In order to overcome that problem, solutions *(d)* and *(e)*, explicitly mention the "Unknown" value and definitely require a selection, preventing the validation of the form with a missing value for that item. Thus, they are the most informative solutions. Of course, they are also the most time-demanding ones, since in the case of an "Unknown" value, the operator cannot intentionally skip the item but he/she is forced to click once for solution *(d)*, or even twice for *(e)*. In summary, in order to select the most appropriate data input mode, we should know whether an item could be "Unknown" at the data entry time, and, in that case, assess whether or not it is important to distinguish between "No" and "Unknown" options.

Another frequent user complaint is that data entry is too time-consuming. A common solution to streamline data entry consists of pre-instantiating all items with informative default values (e.g., choosing the most frequent value as the default). However, in that case, users could pay less attention to the input fields and stay with the default values even when they are not the correct ones. In order to compensate for that bias, we decided to use the "missing value" as default.

Other considerations come from the GUI layout. Figure 1 shows solutions with a more or less compact layout to be chosen also according to the size of the device screen (PC vs. tablets or mobiles).

Now let's consider numerical items. As already mentioned, in that case the input can be checked against the range of admissible values. However, when those values fall into a relatively small finite set (e.g., in the Rankin scale the value is an integer between 0 and 6), we can provide a dropdown menu constraining the user to select just one of those items. This has two advantages; first, it avoids further checks, and second, it recalls the set of values for the user, letting him/her better realize the meaning of that item.

As a last example, consider the encoding of values. This is a highly recommended practice to achieve a uniform terminology across different databases, in view of data aggregation. However, browsing a large terminology server (ICD9-CM and ICD10 list more than 10,000 pathologies and interventions) is far from an easy or quick task [29]. When dealing with those classifications, the usual approaches, such as those encompassing an *autocompletion* feature, prove to be inadequate either because they retrieve too many codes or because the user does not properly enter the leading part of the term (e.g., he/she is familiar with a synonym) and nothing is retrieved as a consequence. A possible solution is to show a dropdown menu that includes the most frequent items only, and let the user navigate across the entire terminology only when the required item is not found. The system may also be instructed to progressively learn the actual frequency of the chosen items, and therefore automatically update the menu.

When it is not possible to enumerate all the different values of an item, one may adopt a mixed approach by enumerating the most frequent choices as encoded values, and then offering a last choice labeled "Other" that, when selected, unblocks a free-text input. This approach should allow to keep track of rare and non-standard choices, but it entails a risk. In fact, when the user sees the "Other" choice, he/she may be erroneously tricked into believing that for better answering the question, he/she should go for it even though, with respect to the model adopted by the study designer, he/she should opt for one of the predefined options. This mixed approach was exactly the one adopted in the first release of the SUN Registry, and during the re-engineering phase, we examined the free text entered with the aim of identifying possible missing items and completing the choices. Interestingly, about 50% of the "Other" choices could well fit the already existing menu items, indicating that the user did not understand the meaning of those items. The remaining, informative, "Other" specifications were used to add new items to the dropdown menus. After two such iterations, each one lasting for one month, "Other" was not used anymore, so it was definitely dropped from the menu.

In summary, as it transpires from all the examples reported above, the most appropriate mode for entering data emerges as the result of a compromise among an easy and quick use of the system, the users' preferences, and the sought quality level for data completeness and consistency.

3. RESULTS

Considering all the problems emerged while working with the first version of the SUN Registry, we decided to re-engineer it from the ground up, and results are described in the following paragraphs.

3.1. The Sun Registry Architecture

Figure 2 shows the functional architecture of the SUN Registry after the re-engineering process. The registry has been deployed as a web application combining the advantages of a central management of the data repository with the possibility of easily collecting data. The same technology also supports the stroke community through a forum where users may upload/download SUN-related documents and update the profiles of their centers. It is able to automatically calculate and disseminate stroke indicators that are to be interpreted also in light of those profiles. Indicators concern both the stroke care process and its outcomes, and have been agreed-on during consensus meetings with medical experts. The participating centers may obtain automatic reports about their patients and visualize their own specific indicators as well as the overall ones resulting by pooling data from all centers. The coordinator may visualize reports for any patient and indicators for any center, in addition to the overall ones. The database is kept and managed at the coordinating center, but data export and statistics are available at any time to every user within the SUN.

3.2. Data-entry Checks and User Notifications

In order to prevent data entry errors, a special emphasis during the re-engineering has been devoted to the development of the GUI, taking into account the importance of both

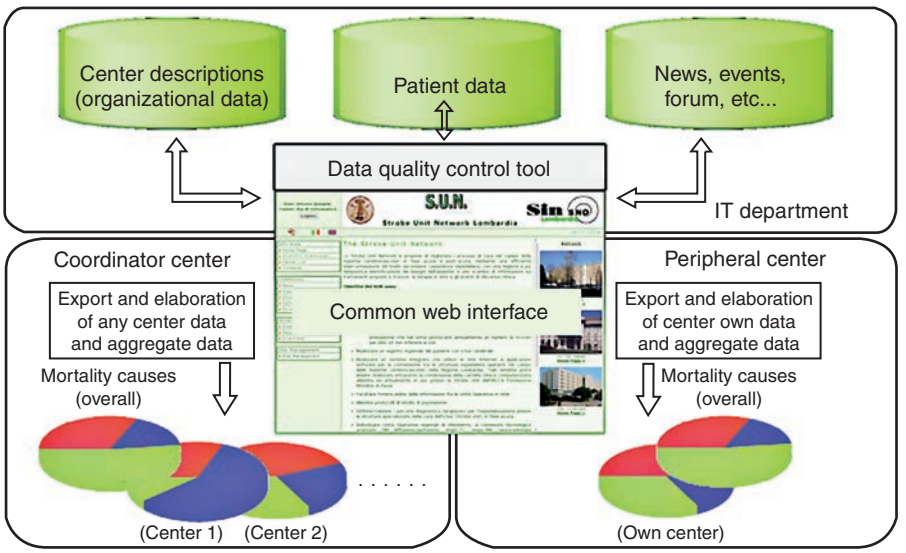


Figure 2. The functional architecture of the new release of the SUN Registry.

informing users about errors, and avoiding confusion and frustration when those errors are detected or suspected. Thus, we classified them *as definite errors, probable errors* and missing values. A comprehensive set of consistency checks and constraints have been based on that error model for generating notifications, as illustrated below.

3.2.1. Temporal Constraints Satisfaction

Temporal constraints are very important because sensible statistics on a specific event may be accomplished only when there is enough information about previous related events. Even more, process mining relies on a correct temporal sequence of events.

First of all, a time sequence is imposed so that events are entered in a pre-defined order: *Emergency*, SU *Admission, Discharge*, and *Follow-up*. This is achieved by accessing subsequent forms only after the preceding ones have been properly filled out. To further limit the error possibility, precedence rules have been implemented among several couples of dates, raising a *definite error* when they are violated, as indicated by the message shown in the bottom line of Figure 3. Additional checks require assuming a maximum time interval between events. For example, the longest interval between the symptom onset and the arrival at the emergency room is assumed to be 60 days, while the maximum stay in the SU is assumed to be 100 days. These thresholds, decided through the medical experts' consensus during the phase of requirement analysis, may be overridden when unexpected exceptions intervene in a patient's clinical path. Therefore, their violation is ranked as probable error and it generates only a warning. A collection of those warnings is then used for updating the constraints during the periodic evaluations of the system.

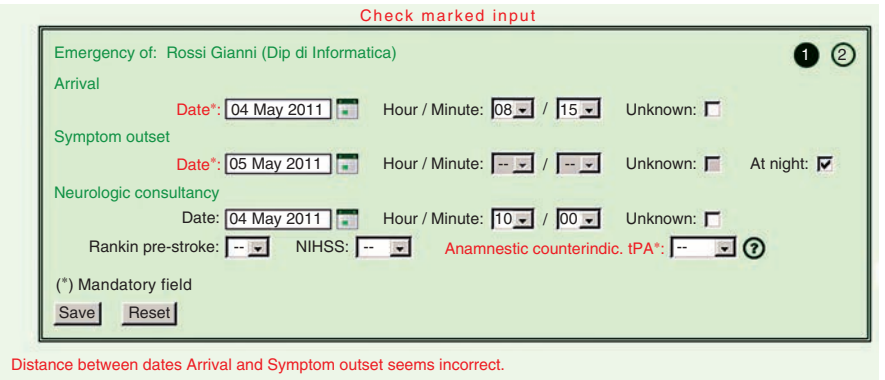


Figure 3. Data entry checks. Erroneous or missing data are actually shown in red. At the bottom, an explanation is provided. Stars indicate mandatory fields. The question mark near the contraindication for tPA (the thrombolytic drug treatment) calls the guideline page with all those possible contraindications.

3.2.2 . Selective Activation/Deactivation of Input Items

In order to acquire a rich data set for statistical purposes, each event features a wide set of input fields. Sometimes, depending on the previous data entered, some of those input fields are not applicable. When this happens, they are automatically disabled in order to inform the user about the situation and avoid acquiring inconsistent data. For example, if the stroke onset occurred at night when the patient was sleeping, only the date is mandatory, since in that case the time will be unknown. This is shown in Figure 3 where the controls for entering the onset time for stroke have been disabled.

3.2.3 . Mandatory Fields

Mandatory fields should be kept at a minimum, because when such data are unavailable, they prevent further progress in the data entry process by raising *a missing value* error. In the SUN Registry, those are limited to some demographic data, which are required to validate the entire case; timing of symptoms onset and hospital arrival, which are necessary for identifying the start of the process; contraindication to thrombolysis, since it represents the most important treatment for ischemic stroke. Moreover, we also implemented *conditionally mandatory* data entry checks, exploiting the above mentioned selective activation of menu items. For example, if the item “antiplatelet treatment” is set to “Yes”, then the specific antiplatelet drug must be entered.

Minimizing the number of mandatory fields increases the flexibility but, on the other hand, also the risk of a partial completion of the forms. Therefore, we provided the coordinator with a means to check the data completeness, i.e., he/she can automatically obtain the missing value rates for every data item.

3.3. Help On-Line

An on-line guide has been designed to help the users in better understanding the meaning of the information to be entered. Its contents have been structured at three different levels, namely context, input form, and single item.

Context-related material explains the rationale for collecting a set of data which can be entered through one or more forms. For example, SUN registry data belong to the "stroke" context, that can be further refined into narrower contexts such as *emergency*, *SU admission, discharge and follow-up*. Form-specific help instructs the user about the intents and the features of each input form, relating those to the specific context they belong to. Finally, item-related help addresses any possible ambiguity about the meaning of a specific form item. Some help sentences are the following:

- At the context level:
  1) *"Emergency data are not necessarily referred to an emergency room, but to any environment where the emergency is managed (e.g., stroke could develop in a hospital ward, in which case the emergency may be managed in the ward itself)";*
  2) *"Follow-up may be performed during a control visit, or by telephone"*.
- At the data-entry form level:
  1) *"This is one of the multiple forms in this section: the "save" button will check if all the mandatory items in the previous forms have been properly filled out";*
  2) *"This form collects data about the examinations performed during the patient's stay in the SU. All timestamps must be referred to the SU admission time. That's why, as a reminder, you can see that admission time on top of the form"*.
- At the single item level:
  1) *"This item (e.g., the thrombolytic treatment) is mandatory because it represents a crucial, life-saving treatment";*
  2) *"Please note that the accomplishment state for the following diagnostic tests has been assigned the default value of 'Unknown' which is different from 'Not performed'. Please go through the drop-down menu and select a value if the state is known"*.

We often added explanatory labels for date and time data, which is very important as some data may be associated to different timestamps. For example, a diagnostic test may have the time of the sample (e.g., the time at which a blood sample has been drawn from the patient); the time of the measurement (e.g., the time at which a blood sample has been processed by the lab); the time of the referral (e.g., the time at which the results have been reported to the patient's physician). By interviewing the data entry personnel, we realized that different interpretations actually did occur.

3.3.1. Knowledge-based Help

In addition to plain explanations used by the on-line help, we also explicitly represented any relationship, based on medical knowledge, useful for detecting treacherous data entry errors, such as:

- Anticoagulant Treatment *treats* Ischemic Stroke
- Fresh Frozen Plasma *treats* Intracranial Haemorrhages
- Warfarin *is-a* Anticoagulant Treatment
- Atrial Fibrillation *is-a-risk-factor-for* Cardioembolic Ischemic Stroke
- Rankin=6 *implies* Death (and vice-versa)

These relationships have been organized into a semantic network, and are exploited to generate warnings whenever the actual data violate them. This may occur, for example, if Fresh Frozen Plasma is selected to treat a patient with Ischemic Stroke, or if a patient with "Atrial Fibrillation" is diagnosed with "Atherothrombotic Stroke" (which is possible although unlikely). The semantic network has been designed together with medical experts and then represented implicitly in the SUN Registry in terms of validation controls written with the underlying implementation language of the system (i.e., Java). The visual representation of the same semantic network is used to better explain the meaning of a data item. To illustrate this functionality, let us refer to tPA which is a crucial treatment for ischemic stroke and deserves a great attention to possible situations which are not compatible with its administration. In the GUI, there is more than one data item addressing contraindications to tPA, with different items belonging to different parts of the same forms. For example, how can we help the non-medical data-entry personnel in properly interpreting the label "anamnestic contraindications to tPA" in the top left part of Figure 4, which shows a section of the *emergency form?* The link displaying the semantic network allows the user to realize that this particular item is related to the patient *clinical history*, thus addressing him/her to the correct data source, namely clinical history documents. Similarly, the label in the top right of Figure 4, concerning the tests accomplished in the emergency room, allows to realize that the "Contraindication to tPA" menu item refers to some threshold values for blood glucose and platelet, that can be matched with values in the biochemical analysis reports.

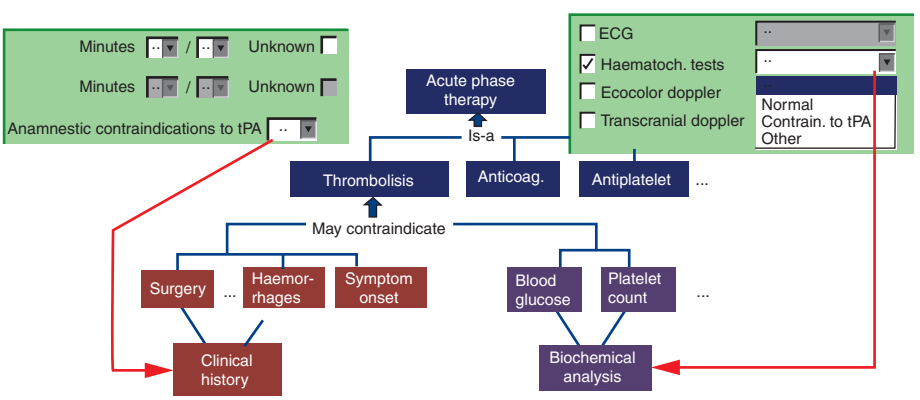


Figure 4. Correlating data items in the GUI with the underlying medical knowledge: a portion of the semantic network.

Table 1. Errors and missing data detected among records of the first release of the SUN Registry. All the numbers reduced to zero after the reengineering.

Error type	Number	Percentage
Violation of temporal constraints among dates	396	7.5%
Stroke assessment scales out of range	14	0.2%
Inconsistency among values of stroke scales*	60	1%
Missing stroke scales at discharge	201	3%
Missing stroke scales at follow-up	1028	22.3%
Missing onset-arrival time	3141	42.9%
Inconsistency/incompleteness with death information	50	0.7%

*Examples: Rankin scale = 6 means death. NIHSS = 1 indicates mild impairment.

3.4. Improvements in Data Acquisition and Quality

In order to assess the improvements achieved in terms of both the quality of collected data and the easiness of use, we performed a comparative analysis of the two versions of the SUN Registry before and after the re-engineering process, respectively. The basis for this comparison is represented by Table 1 which reports the percentages of the records affected by different types of errors or missing data in the first release of the SUN Registry. As already mentioned above, errors and missing data in EMRs and DRs severely impair their reliability in supporting patient management and their usability for subsequent sound statistics. In stroke-related applications, for example, missing the symptom onset time or the arrival time at the hospital (see the last-but-one row in Table 1) makes it impossible to assess a patient's eligibility for thrombolysis, that is a crucial treatment for stroke.

Since the release of the new version of the SUN Registry in July 2009, about 15,000 more patients' data have been collected. As a confirmation of the correct implementation of data quality control, we issued the same queries that produced the results indicated in Table 1, with the former version of the registry, against the newly collected patients. All the counts fell to zero witnessing that the re-engineering process succeeded and all those errors were completely fixed. We also shadowed the data-entry personnel during the first month of their activity with the new release in order to assess the perception of the user experience, fix bugs, and measure times. The average time for entering a complete case was 15 ± 3 min which is one minute longer on the average than the time required before the re-engineering process. However, given the complete elimination of the most frequent data entry errors that required additional time for their discovery in the first release, this has been considered as a success by the coordinator.

3.5. Enhancing Networking Among Centers

As we stated in the Introduction, one of the aims of the SUN Registry is to improve the stroke care process in all the involved centers. To achieve this goal, the re-engineering also took into account some important networking functionalities, which allow every center to compare its own performance with the overall one, to share news and documents, and to monitor patients' enrolment.

To properly calculate performance indicators and provide a reliable comparison among centers, it is important to know the resources available at every center. SU heads enter this information through the SUN Registry website. Each data set is entered with the reference date, and whenever a center changes its features, a new record is created with the corresponding date. This produces a historical archive allowing, in a retrospective analysis, to correlate processes of care to the resources available at that time. SU profiling sheets have been filled out by 31 centers out of 39 (79.5%), and the completeness of the information supplied was very good (2.8% missing data). Results such as those shown in Table 2 are automatically made available to the SUN coordinator.

Moreover, through the SUN website, users may access a Reports area where the indicators for the center they belong to can be automatically computed after selecting a patients population. The selection is based on a set of attributes including age, type of stroke, severity, and, for the coordinator, also the resource level of SUs. Using this targeting functionality, the coordinator, who may analyze data from all the centers, can compare SUs' performances by adjusting for the case-mix. As an example, Figure 5 reports the variability among centers for some indicators. The pies show the mortality at discharge for patients with intracranial hemorrhages (ICH). The three additional rows report the compliance with stroke guidelines, which recommend (i) using scales, such as the NIHSS, to assess the patient status, (ii) prescribing statins to ischemic patients, and (iii) starting rehabilitation as soon as possible. The figures in parentheses show the patients' adherence to statins prescription, as observed at follow-up. Those results are useful to identify the steps of the care process that need to be improved and are also available on a per-center basis.

3.5.1. Patient Enrolment

After the re-engineering, the new version of the SUN Registry was endowed with a functionality which helped the coordinator in monitoring several aspects of the data entry process. First of all, the coordinator can watch the temporal trend of the patient enrolment and judge the enrolment rate according to the expected number of cases for the current

Table 2. Statistics of diagnostic test availability within the SUs in the SUN registry

Test	Availability	N (%)
Transesophagea	h24*	4 (13.0 %)
Echocardiogram	Only some days per week	10 (32.2 %)
	Only during the day	17 (54.8 %)
Transthoracic	h24*	18 (58.1 %)
Echocardiogram	Only some days per week	3 (9.7 %)
	Only during the day	10 (32.2 %)
EcoDoppler TSA	h24*	18 (58.1 %)
	Only during the day	13 (41.9 %)
Cerebral CT	h24*	31 (100 %)

24 hours a day

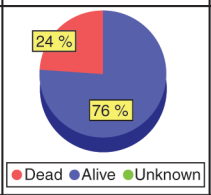
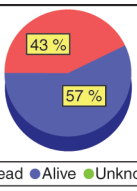
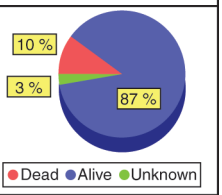
	Overall	Best case	Worst case
Outcome at discharge for ICH patients			
NIHSS measurement at admission	66.8%	98.2%	4.8%
Statins prescriptions (adherence at follow-up) for ischemic patients	39.6% (73.0%)	78.0% (80.6%)	16.0% (38.3%)
Early rehabiliation	49.0%	73.5%	20.7%

Figure 5. Variability among centers: outcome at discharge for ICH patients (pies), and compliance with three important recommendations of the stroke guidelines. "Best" and "Worst" cases refer to the SUs with the best and worst outcome for the specific finding reported in each table row.

year. The latter is estimated on the basis of the patients' discharge reports that centers provided during the previous year to the regional healthcare management service. The SUN coordinator may use those enrolment figures during the users group meetings, showing both the *white* and the *black* lists of centers. We measured the effect of such a comparison during a meeting at the end of 2011. The average number of enrolled cases per month increased from 720 (average over 3 months before the meeting) to 812 (average over 3 months after the meeting), an increase of about 13%.

3.6. Data Analysis

A detailed description of the results emerged through the analysis of the SUN Registry data is beyond the scope of this paper. However, the most interesting results are summarized below, referring the reader to the literature for further details. The aims of these analyses were the following:

- finding the highest non-compliance rate with respect to the stroke guidelines recommendations;
- detecting the factors influencing the physicians' behavior causing such non-compliance;
- finding correlation between the compliance with guidelines and health outcomes.

Our team [30] used machine learning techniques, namely classification trees, for identifying both the guidelines recommendations and the organizational variables (i.e., human and technical resources) that are most relevant to the health outcome, coded as good (Rankin scale > 1) or bad (Rankin scale ≤ 1). Interestingly, the analysis showed that some recommendations had a negative impact on health. In particular, they were concerned with the use of graded compression stockings to prevent deep venous

thrombosis and lowering of arterial blood pressure. About the first issue, our findings have been confirmed by other independent studies [31], and the updated guideline does not contain that recommendation anymore. About the second one, we have individuated a subpopulation, namely the elderly, for whom low blood pressure was associated to a worst outcome at discharge.

We predicted the chance to be treated with thrombolysis [32]. The factors related to a greater chance were severe stroke, neurological evaluation performed before the neuroimaging, admission to an emergency department linked to a SU, and the 24-hour availability of a neurologist in the emergency department.

Eventually, the large amount of data collected is an excellent test bench for exploring new algorithms for data and process mining. We proposed a new definition of distance to assess the similarity among care processes produced by different SUs [33]. We called it "taxonomic distance", since it takes into account a taxonomy of all the medical actions that can be recorded by the registry. In a taxonomical hierarchy, actions having the same goals are closer than actions having different goals. In the mapping procedure that compares two care processes, this allows to compute a more sensible distance among them. The algorithm has been validated in extracting the process of a specific SU, taking this as reference model, and discovering the SUs that behaved similarly. Compared to algorithms that use standard measures of distance, this new one performs better in retrieving SUs of the same level of the reference unit.

4. DISCUSSION

Stroke registries exist in several countries, and show different characteristics according to the goals they pursue. The Innsbruck stroke registry in Germany [34] was built with the main purpose of assessing the safety and the effectiveness of thrombolysis, while the China Interventional Stroke Registry [35] was developed to assess angioplastic and stenting procedures. Thus, they are oriented to specific problems. On the contrary, other registries have a wider purpose, in that, similarly to the SUN Registry, they are aimed at tracking and improving the overall process of stroke care, monitoring a variable number of quality-of-care indicators. Examples are given by the Paul Coverdell National Acute Stroke Registry Surveillance founded by the Center for Disease Control (CDC) in the US [36], the Registry of the Canadian Stroke Network [37-38], and the China National Stroke Registry [39]. Data collected are comparable to the SUN Registry both in their amount and type. The Korean Stroke Registry, launched in 2002 with analogous purposes, has been recently exploited also to evaluate national time trends in the clinical presentation of stroke [40]. The Indian registry instead has the aim of explaining the increasing incidence of stroke in the country and is particularly focused on risk factors [41].

All those registries share the common goal of sending feedback to the participating centers in order to improve stroke care. The majority of them address the issue of data quality. While they all implement simple checks, none of them mentions more sophisticated real-time data entry controls. As a consequence, *a-posteriori* checks and corrective actions are adopted. For example, after each quarterly data submission, the CDC provides state registry programs with individual reports on missing, invalid, and

questionable data. A procedure common to several registries is to select a random sample of patients' records, and re-abstract them from the original data sources in order to assess the accuracy of the former abstractions. Finally, most of the coordinating centers organize training sessions for data-entry personnel. Nonetheless, as also pointed out by Jung [40], these controls may be inadequate to preserve a good data quality, for a variety of reasons including changes of individual hospital policies and turnover of dedicated staff. Wang et al. [39] report on some variables with more than 10% of missing values, causing their exclusion from statistical analyses. The above observations highlight the importance of implementing a tighter real-time data entry control, also considering the costs of the *a-posteriori* check procedures that require lot of time and dedicated medical experts. We do not claim that human experts' checks can be completely replaced by automatic controls, but they could be significantly reduced.

Despite the encouraging results obtained with the quality of the SUN Registry data, our study still suffers from some limitations. A first issue is given by the lack of a formal stroke ontology for explicitly representing the registry domain. As a matter of fact, data quality checks based on medical knowledge have been implemented ad hoc, through a set of IF-THEN rules derived by the relationships shown in section 3.3.1. This approach could be greatly enhanced by framing this knowledge into a stroke ontology. Similar efforts are described in the literature, not only within the healthcare informatics area. For example, Cannon et al. [42] developed a tool for generating a GUI from a flowering plants ontology to allow plant taxonomists describing the specimens used during the classification process. Wang et al. [43] describe how an ontology can improve data entry in mobile applications for environmental protection. Ontologies are widely used in medical informatics, but mainly for data integration and interoperability, data and text mining, and translational bioinformatics [44-48]. In spite of those efforts, a recent review [49] calls for more practical applications since, despite their potential, "Ontologies and semantic integration methods are emergent with limited evidence-base for their implementation". A specific application for medical data entry is described that created an ontology of the immunological system whose main purpose is to allow users entering data at different hierarchical levels, according to the available data detail [50]. OnWARD, an ontology-driven web-based tool is reported for data collection in multicenter clinical trials [51]. In line with the mentioned experiences, a future development of the SUN Registry will include the implementation of a stroke ontology. However, different from the above studies, our ontology will also include the data sources (electronic or paper-based). This is useful because, unfortunately, double data-entry is still required in many applications.

A second issue causes some problems still lingering in the GUI. In particular, some data-entry forms are too time-consuming according to the users' complaints. A possible compromise between compilation time and the quality of data, could be achieved when there is a set of items belonging to the same logical section of a form. For example, in the SUN Registry, we have 12 stroke risk factors that may take "Yes" or "No" values. As pointed out in Section 3.3, they are initialized with a "missing value" and the user is then forced to enter "Yes" or "No" going through the entire set. However, once the user fills out two values (showing that the user is paying attention to those items), the

GUI could allow the user to automatically set all the remaining ones to "No". This represents an advantage because the median number of risk factors per patient is 2 (first and third quartiles equal to 1 and 3, respectively, as it emerges from our data). This is an example of how the best data entry policy may be assessed only when an adequate amount of data are available to perform some basic statistics.

A third improvement could be the automatic selection of the patients to be interviewed for the follow-up three months after their discharge. This selection is currently performed by an off-line calculation over the data exported from the registry on a weekly basis. That task could be further sped-up by an external connection with the mortality database managed by the Regional Healthcare System (similarly to the Korean registry, that is connected to a national death certificate system).

The latter issue raises a wider discussion about the opportunity of interconnecting registries both among them and with other electronic data sources. As a matter of fact, a patient could be affected by more than one disease, resulting in enrollment in two or more registries. In our domain, for example, stroke shares risk factors with all the other cardiovascular diseases; thus, a stroke patient is also likely to suffer from diabetes, hypertension, etc., and he/she is also at risk for myocardial infarction. If different registries exist for all or some of these pathologies, data and process mining would greatly benefit from their integration. At the same time, integration poses technological as well as organizational problems, and sharing patient data among different institutions is also related to the important issue of preserving data privacy, usually accomplished through anonymization procedures. This task is greatly simplified when the information is entered into a registry only once for the same patient, and the registry will serve purely epidemiological purposes. In fact, in that case, there is no need to keep a personal reference to the patient. However, for the SUN Registry, acquiring data was inherently a two-step process. First, the information about the patient stay at the hospital was entered. That information needs to be recalled after 3 months in order to be complemented with the follow-up data. We considered several alternatives for the patient identification problem. The first one envisioned generating a hash code starting from the patient identification data (i.e., the taxpayer ID). However, this approach was rejected since it required to supply each center with a separate hash code generator making the whole task of using the registry more complex. Moreover, centers couldn't always use the same data for identifying patients (e.g., some immigrants do not have a taxpayer ID). Thus, as other DR developers [52], we resolved to use a mixed approach. Since the SUN Registry assigns its own unique identifier to each new patient, the participating centers were allowed (albeit not forced) to use this anonymous code for recalling patients. In that case, they would not need to enter any other personal identification data. This approach has been mostly used by centers equipped with an internal EMR since they could add a note to each patient's record to keep track of the ID assigned by the SUN Registry. Other centers enter just the patient initials and used those along with the date of birth, which is a required item, for recalling patients. In case of homonyms, they use a nickname to disambiguate. Finally, other centers used the SUN Registry patient record as a full identification record, including taxpayer code.

As it is clear from the above discussion, this is a sensible topic which has no straightforward solution, since registries are often managed by research institutions which are not directly related to the centralized ones officially appointed for collecting patient data on a regional or national basis. With reference to our experience, the SUN Registry has been claimed by the Regional Healthcare System for its incorporation into the national electronic health dossier only after 5 years of service directly managed by the coordinator institution. We hope that this will facilitate its integration with other data sources in the near future.

5. CONCLUSION

This paper described the re-engineering of a disease registry for collecting data about stroke patients. The work originated from an analysis of the previous repository which showed severe limitations due to errors and missing data after a two-year experience as a source of reliable statistics. Possible causes were erroneous interpretations of the data labels in the GUI, wrong transcription, and access to the wrong data sources. In fact, the registry was filled out, after the patient's discharge, by collecting data from different sources, both electronic and paper-based. To overcome those problems, we carried out an in-depth requirement analysis involving medical domain experts to develop a knowledge-based interface allowing data entry personnel with different skills and background to share the same cognitive data model. After the re-engineering process, we monitored the registry data for a period of three years during which we proved that such process helped decreasing data-entry errors.

Although this paper deals with a registry related to a stroke application, we believe that the analysis on data quality issues and the proposed solutions for reducing errors during data-entry are also applicable to health information systems targeted at different domains requiring a close interaction with users belonging to multiple centers. In that case, we hope that the developers could benefit from this paper in better approaching and accomplishing the requirement analysis and the design procedures of their systems.

ACKNOWLEDGEMENTS

This work was partially funded by the Italian Ministry of Health. We thank all the participants to the SUN network for providing stroke patients data.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to report.

REFERENCES

[1] The Committee on Quality Health Care in America. Crossing the quality chasm: a new health system for the 21st century. National Academy Press, ISBN 0-309-07280-8 , 2001. http://www.nap.edu/catalog.php?record_id=10027. Accessed Dec 25, 2013.

[2] Schmittdiel J, Bodenheimer T, Solomon NA, Gillies RR, Shortell SM. Brief report: The prevalence and use of chronic disease registries in physician organizations. A national survey. *Journal of General Internal Medicine*, 2005, 20(9):855–8.

[3] Colias M. Disease registries. *Hospitals and Health Networks*, 2005, 79(2):62-4, 66–8, 2.

[4] Arts DGT, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: A literature review, case study, and generic framework. *Journal of the American Medical Informatics Association*, 2009, 9(6):600–611.

[5] Goldberg SI, Niemierko A, and Turchin A. Analysis of Data Errors in Clinical Research Databases. Proceedings of the AMIA Annual Symposium, 2008, 242–246.

[6] Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of electronic health records in U.S. hospitals. *New England Journal of Medicine*, 2009, 360(16):1628–1638.

[7] Yan H, Gardner R, Baier R. Beyond the focus group: understanding physicians' barriers to electronic medical records. Joint Commission Journal on Quality and Patient Safety, 2012, 38(4):184–191.

[8] Sleeth-Keppler D. White Paper: Garbage In, Garbage Out. Survey-Question Design. A free report.VALS™ Strategic Business Insights http://www.strategicbusinessinsights.com/ vals/free/2010-02whtpprsurveydes.pdf. Accessed Dec 25, 2013.

[9] Hartzband P, Groopman J. Off the record—avoiding the pitfalls of going electronic. *New England Journal of Medicine,* 2008, 358(16):1656–8.

[10] Peute LW, Driest KF, Marcilly R, Bras Da Costa S, Beuscart-Zephir MC, Jaspers MW. A Framework for reporting on human factor/usability studies of health information technologies. Studies on Health Technologies and Informatics, IOS Press, 2013, 194:54–60.

[11] Di Carlo A. Human and economic burden of stroke. *Age and Ageing*, 2009, 38(1):4–5.

[12] Stroke Unit Trialists' Collaboration. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews*, 2007, 17(4):CD000197.

[13] Langhorne P, Lewsey JD, Jhund PS, Gillies M, Chalmers JW, Redpath A, Briggs A, Walters M, Capewell S, McMurray JJ, MacIntyre K. Estimating the impact of stroke unit care in a whole population: an epidemiological study using routine data. *Journal of Neurology Neurosurgery and Psychiatry*, 2010, 81(12):1301–5.

[14] Bersano A, Candelise L, Sterzi R, Micieli G, Gattinoni M, Morabito A. Stroke Unit care in Italy. Results from PROSIT (Project on Stroke Services in Italy). A nationwide study. *Neurological Sciences*, 2006, 27(5):332–9.

[15] Candelise L, Gattinoni M, Bersano A, Micieli G, Sterzi R, Morabito A. Stroke-unit care for acute stroke patients: an observational follow-up study. *The Lancet*, 2007, 369(9558):299–305.

[16] Cavallini A, Micieli, G. Lombardia stroke unit network project. *Neurological Sciences*, 2006, 27(Suppl 3):S268–S272.

[17] Gensini GF, Dilaghi B, Zaninelli A. Italian SPREAD Guidelines: from past to future. *Neurological Sciences*, 2006, 27(Suppl 3):S254–7.

[18] Micieli G, Cavallini A, Quaglini S, Fontana G. Due' M. The Lombardia Stroke Unit Registry: 1-year experience of a web-based hospital stroke registry. *Neurological Sciences*, 2010, 31(5):555–564.

[19] Rahbar MH, Gonzales NR, Ardjomand-Hessabi M, Tahanan A, Sline MR, Peng H, Pandurengan R, Vahidy FS, Tanksley JD, Delano AA, Malazarte RM, Choi EE, Savitz SI, Grotta JC. The University of Texas Houston Stroke Registry (UTHSR): implementation of enhanced data quality assurance procedures improves data quality. *BMC Neurology*, 2013, 13:61.

[20] Boonstra, A, Broekhuis, H. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC Health Services Research*, 2010, 10:231.

[21] Dornan S, Murray FE, White G, McGilchrist MM, Evans JM, McDevitt DG, MacDonald TM. An audit of the accuracy of upper gastrointestinal diagnoses in Scottish Morbidity Records 1 data in Tayside. *Health Bulletin*, 53(5):274–9.

[22] Warsi AA, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. *European Journal of Surgical Oncology*, 2002, 28(8):850–6.

[23] Schaff HV, Brown ML, Lenoch JR. Data entry and data accuracy. *Journal of Thoracic Cardiovascular Surgery*, 2010, 140(5):960–1.

[24] Shrawankar U, Thakare V, Parameters optimization for improving ASR performance in adverse real world noisy environmental conditions. *International Journal of Human Computer Interaction*, 2012, 3(3):58–70.

[25] Panzarasa S, Quaglini S, Pessina M, Cavallini A, Micieli G. GIFT: a tool for generating free text reports from encoded data. *Conf Proc IFMBE, 11th Mediterranean Conference on Medical and Biomedical Engineering and Computing*, 2007, 16:152–156.

[26] Los RK, Roukema J, van Ginneken AM, de Wilde M, van der Lei J. Are structured data structured identically? Investigating the uniformity of pediatric patient data recorded using OpenSDE. *Methods of Information in Medicine*, 2005, 44(5):631–8.

[27] Quaglini S, Ciccarese P, Micieli G, Cavallini A. Non-compliance with guidelines: motivations and consequences in a case study. *Conf Proc CGP 2004 Symposium on Computerized Guidelines and Protocols*. Studies in health technology and informatics, IOS Press, 2004, 101:75-87.

[28] Quaglini S. Compliance with clinical practice guidelines. In: Annette Ten Teije, Silvia Miksch, Peter Lucas (eds.), *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends,* Studies in health technology and informatics, IOS Press, 2008, 139:160–179.

[29] Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, Frost JP, Kuperman GJ. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. *International Journal of Medical Informatics*, 2003, 72(1–3):17–28.

[30] Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, Stefanelli M. Data mining techniques for analyzing stroke care processes. *Conf Proc MEDINFO*. Studies in Health Technology and Informatics, 2010, 160(2):939–943.

[31] Dennis M, Sandercock PA, Reid J, Graham C, Murray G, Venables G, Rudd A, Bowler G. Effectiveness of thigh-length graduated compression stockings to reduce the risk of deep vein thrombosis after stroke (CLOTS trial 1): a multicentre, randomised controlled trial. *The Lancet*, 2009, 373(9679):1958–1965.

[32] Cavallini A, Tartara E, Marcheselli S, Agostoni E, Quaglini S, Micieli G. Improving thrombolysis for acute ischemic stroke in Lombardia stroke centers. *Neurological Sciences*, 2013, 34(7):1227–1233.

[33] Montani S, Leonardi G, Quaglini S, Cavallini A, Micieli G. Mining and Retrieving Medical Processes to Assess the Quality of Care. *Conf Proc ICCBR*, 2013:233–240.

[34] Matosevi? B, Zangerle A, Furtner M, Knoflach M, Werner P, Prantl B, Wille G, Illmer A, Mair A, Spiegel M, Schmidauer C, Sojer M, Muigg A, Willeit J, Kiechl S. Implementation of thrombolysis in acute stroke—10-year results of the Innsbruck stroke registry. *Wiener Klinische Wochenschrift*, 2009, 121(23-24):750–6.

[35] Liu X, Xiong Y, Zhou Z, Niu G, Wang W, Xiao G, Lin M, Leung TW, Liu D, Liu W, Fan X, Yin Q, Zhu W, Ma M, Zhang R, Xu G. China interventional stroke registry: rationale and study design. *Cerebrovascular Diseases*, 2013, 35(4):349–354.

[36] George MG, Tong X, McGruder H, Yoon P, Rosamond W, Winquist A, Hinchey J, Wall HK, Pandey DK. Paul Coverdell National Acute Stroke Registry Surveillance - four states, 2005-2007. *MMWR Surveillance Summaries*, 2009, 58(7):1–23.

[37] Lindsay MP, Kapral MK, Gladstone D, Holloway R, Tu JV, Laupacis A, Grimshaw JM. The Canadian Stroke Quality of Care Study: establishing indicators for optimal acute stroke care. *Canadian Medical Association Journal*, 2005, 172(3):363–5.

[38] Kapral MK, Hall R, Stamplecoski M, Meyer S, Asllani E, Fang J, Richards J, O'Callaghan C, Silver FL. *Registry of the Canadian Stroke Network: Report on the 2008/09 Ontario Stroke Audit*. Toronto: Institute for Clinical Evaluative Sciences, 2011.

[39] Wang Y, Cui L, Ji X, Dong Q, Zeng J, Wang Y, Zhou Y, Zhao X, Wang C, Liu L, Nguyen-Huynh MN, Claiborne Johnston S, Wong L, Li H. The China national stroke registry for patients with acute

cerebrovascular events: design, rationale, and baseline patient characteristics. *International Journal of Stroke*, 2011, 6(4):355–361.

[40] Jung KH, Lee SH, Kim BJ, Yu KH, Hong KS, Lee BC, Roh JK. Secular trends in ischemic stroke characteristics in a rapidly developed country: results from the Korean Stroke Registry Study (secular trends in Korean stroke). *Circulation, Cardiovascular Quality and Outcomes*, 2012, 5(3):327–334.

[41] Bhaumik S. India launches stroke registry to combat "epidemic". *British Medical Journal,* 2013, 346:f223.

[42] Cannon A, Kennedy JB, Paterson T, Watson M. Ontology-Driven Automated Generation Of Data Entry Interfaces. *Conf Proc 21st British National Conference On Databases*. Lecture Notes in Computer Science, Springer-Verlag, 2004, 3112:150–164.

[43] Wang F, Mäs S, Reinhardt W, Kandawasvika A. Ontology based quality assurance for mobile data acquisition. *Conf Proc 19th international conference on Informatics for Environmental Protection: Networking Environmental Information*. Brno, Czech Republic, 2005:334–341.

[44] Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, 2009, 10(Suppl 2):S1.

[45] Rector A. Knowledge driven software and "fractal tailoring": Ontologies in development environments for clinical systems. *Conf Proc FOIS*, *Formal Ontology in Information Systems*, IOS Press, 2010:17–28.

[46] Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. *Yearbook of Medical Informatics*, 2013,8(1):132–46.

[47] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association,* 2010, 17:124–130.

[48] Falasconi S. Lanzola G, Stefanelli M. Ontology and terminology servers in agent-based Health-care Information Systems. *Methods of Information in Medicine*, 1997, 36(1):30–43.

[49] Liyanage H, Liaw ST, Kuziemsky C, Terry AL, Jones S, Soler JK, de Lusignan S. The Evidence-base for Using Ontologies and Semantic Integration Methodologies to Support Integrated Chronic Disease Management in Primary and Ambulatory Care: Realist Review. *Yearbook of Medical Informatics*, 2013, 8(1):147–154.

[50] Zavaliy T, Nikolski I. Ontology-based information system for collecting electronic medical records data. *Conf Proc TCSET, IEEE Conference on Telecommunications and Computer Science*, Lviv-Slavske, Ukraine, 2010:125.

[51] Tran VA, Johnson N, Redline S, Zhanga GQ. OnWARD: Ontology-driven web-based framework for multi-center clinical studies. *Journal of Biomedical Informatics*, 2011, 44(Suppl 1):S48–S53.

[52] Wahlgren N, Ahmed N, Davalos A, Ford GA, Grond M, Hacke W, Hennerici MG, Kaste M, Kuelkens S, Larrue V, Lees KR, Roine RO, Soinne L, Toni D, Vanhooren G. Thrombolysis with alteplase for acute ischaemic stroke in the Safe Implementation of Thrombolysis in Stroke-Monitoring Study (SITS-MOST): an observational study. *The Lancet,* 2007, 369(9558):275–282.