Paola Berchialla;Cecilia Scarinzi;Silvia Snidero;Yousif Rahim;Dario Gregori. Information Extraction Approaches to Unconventional Data Sources for "Injury Surveillance System": the Case of Newspapers Clippings. JOURNAL OF MEDICAL SYSTEMS. 36 pp: 475-481.
DOI: 10.1007/s10916-010-9492-1

Information Extraction Approaches to Unconventional Data Sources for "Injury Surveillance System": the Case of Newspapers Clippings

Paola Berchialla[1] & Cecilia Scarinzi2 & Silvia Snidero3 & Yousif Rahim4 & Dario Gregori[1]

[1]Department of Public Health and Microbiology,
University of Torino,
Torino, Italy

[2]Department of Statistics and Applied Mathematics D. de Castro, University of Torino,
Torino, Italy

[3]S&A S.r.l.,
Cuneo, Italy

[4]International Society for Violence and Injury Prevention, Stockholm, Norway

Corresponding author
D. Gregori
Department of Environmental Medicine and Public Health, Via Loredan 18,
35121 Padova, Italy
e-mail: dario.gregori@unipd.it

Abstract

Injury Surveillance Systems based on traditional hospital records or clinical data have the advantage of being a well established, highly reliable source of information for making an active surveillance on specific injuries, like choking in children. However, they suffer the drawback of delays in making data available to the analysis, due to inefficiencies in data collection procedures. In this sense, the integration of clinical based registries with unconven-tional data sources like newspaper articles has the advan-tage of making the system more useful for early alerting. Usage of such sources is difficult since information is only available in the form of free natural-language documents rather than structured databases as required by traditional data mining techniques. Information Extraction (IE) addresses the problem of transforming a corpus of textual documents into a more structured database. In this paper, on a corpora of Italian newspapers articles related to choking in children due to ingestion/inhalation of foreign body we compared the performance of three IE algorithms- (a) a classical rule based system which requires a manual annotation of the rules; (ii) a rule based system which allows for the automatic building of rules; (b) a machine learning method based on Support Vector Machine. Although some useful indications are extracted from the newspaper clippings, this approach is at the time far from being routinely implemented for injury surveillance pur-poses.

Keywords Injury surveillance systems · Text analysis · Injury prevention · Public health · Data mining

Introduction

Injury Surveillance Systems are an essential part of an effective Public Health and safety strategy [1]. Common sources of injury surveillance data are traditional clinical derived data, which include hospital discharge records, death certificates [2] and contains demographic information (name, age, gender and town of residence) relating to the victim along with variable degree of details regarding the circumstances surrounding the accident. Data available from hospital discharge records and from mortality data in death certificates are commonly coded using the Interna-tional Classification of Diseases (ICD-9), although this coding system is not the most appropriate for the purposes of describing injuries, lacking of several details on the circumstances of the injury, Nevertheless, the ICD9-CM remains the most widely coding system used in the public health systems in the world. Indeed, when the accident is the result of an external cause, another code (E-code) should be used in addition for describing both the mechanism and the intent of the injury. However, coding of external causes has been considerably less complete for morbidity data, and this has limited the usefulness of sources such as hospital discharge records for injury surveillance [3]. On the other hand, injuries, especially those involving children, are frequently reported to the public by the media in the immediacy of the event with many details about circumstances and causes [4, 5]. The completeness of injury reporting in newspapers is often disputed [4]. Forst [6], comparing newspaper representations of causes of death with actual causes, concluded that causes of death and their risk factors are disproportionately represented and this misrepresentation may contribute to distorted perceptions of health threats. Also [7] concluded that the media may deliver contradictory reports of health risks or may distort information.

However, despite their limitations, newspapers articles have been suggested for use as injury surveillance tool since they may be a useful adjunct to other injury surveillance efforts. Indeed newspapers articles have been proven to be accurate in supplying information about the circumstances of unintentional injuries from fires and drowning as well as demographic information [8, 9].

In order to make newspaper articles an effective injury surveillance tool, they require to be organized and structured for data mining. In this sense, recognizing clinical words becomes a central issue [10, 11]: solutions to it are basically (a) the unfeasible by-hand revision of entire sets of newspapers and (b) the implementation of automated procedures aimed at extracting information. In more technical terms, the second approach means that the issue has to be faced by implementing Information Extraction (IE) techniques.

Information Extraction is the task concerned with identifying predefined types of information stored in natural language texts. The simplest IE technology is the Named Entity Recognition (NER), which is aimed at identifying all the names specified in users-defined lists or throughout user-defined rules. Since creating rules and list of words to be recognized is difficult and time-consuming, over the past years, a number of IE techniques have been developed to automate this task. Successful techniques include statistical methods, such as Hidden Markov Models and probabilistic context-free grammars, or rule-based methods that employ some form of machine learning. One of the most successful machine learning methods employed in IE is the Support Vector Machine (SVM), which is a general supervised machine learning algorithm and has achieved state-of-the-art performance on many classification tasks.

This paper will focus on IE for Public Health Surveil-lance in foreign body (FB) injuries in children. Although it is a rare event, suffocation due to FB is a major cause of death in children aged 0–3 and it is common also in older age up to 14 years. Recent data [12], which are based on hospital discharge records and death certificates, indicate that the estimated number of choking accidents per year in children aged 0–14 is in the European Union of about 50.000. In Italy in the years 1999–2000 the ratio between the hospitalizations number and the mortality rates was approximately a death every 10 hospitalizations (x 100,000 persons). Due to the emotion this news causes, newspaper accounts may provide detailed information (for example on the type of FB responsible for choking or the children supervision) otherwise difficult to obtain from medical records. With the aim to evaluate IE techniques applied to newspaper clippings, three IE methods were tested on Italian newspaper accounts reporting FB injuries in children aged 0–14. In particular, the IE techniques considered were: (a) NER approach, which requires the definition of list of names throughout to look at in order to recognized key words; (b) a rule based system which allows for automatic building of rules; (c) a machine learning method based on SVM. Following a short presentation of the three techniques, results were compared by means of the performance measures (precision, recall and F-measure) with the aim of finding the best performed algorithm. Finally the usage of IE automated techniques as injury surveillance tool was discussed.

Materials and methods

Data collection

A contract was established with a statewide newspaper clipping service to obtain articles on both fatal and non-fatal children injuries due to suffocation. The clipping service provided Italian newspaper articles describing injuries that occurred during 3 years, since January 2003 until September 2006 in Italy. Articles were selected if the terms “suffocation” and “children” (respectively “soffocamento” and “bambini” in the Italian language based searching) appeared in the text. A total of 388 articles were collected by the newspaper clippling agency, which provided the image of the texts in pdf format. Among them, 44 newspaper articles accounting for unintentional FB injuries in children aged 0–14 were selected and turned into editable text using OCR tools. The selection was carried out reading the title and sub-title of the articles, which provided sufficient information for understanding if suffocation was due to an unintentional ingestion of a foreign body.

Thus the rich text format files obtained were checked against the original for integrity. Finally, they were processed by IE software for retrieving relevant information on age and gender of the injured child and the type FB (organic, pebble, marble, etc.) that caused the accident.

Information extraction techniques

A rule based approach to IE and a machine learning system were compared to the NER approach. All the techniques were tested with the aim to extract relevant information (age and gender of the injured child and the FB that caused the accident) from newspaper accounts of the injuries.

Named entity recognition

NER is a task of Information Extraction aimed at identifying and classifying single words or expressions (two or more words) in free text into predefined categories, such as the type of the FB, the gender of the injured child in our case study. In this paper, NER was considered a benchmark for the evaluation of the other IE techniques, since it has been shown to produce near-human performances [13].

NER was carried out using ANNIE (A Nearly-New Information Extraction system), which is a software freely available as part of GATE (General Architecture for Text Engineering) [14], GATE is one of the most popular tools for Natural Language Processing and has been used for many IE projects and is an open source system, under the GNU library license. ANNIE consists in a set of linguistic components which allow identifying the unit of Natural Language, i.e. words and character punctuation. ANNIE requires the definition of a list of words (gazetteers) for each predefined category. For example words "female" and "male" defined the category "gender of the injured child". Several of such lists were filled in. A first one consisted in a broad range of FBs commonly known from literature to be responsible for choking in children. Other lists were forest down for the "gender of the injured child" category and the "age" category. The gender category consisted of words "male" and "female"; the "age" category consisted in a list of numbers (digit and words) from 0–14. The most frequent pronoun (male or female) and the most frequent number encountered in the article were assigned respectively to the gender and the age of the injured child.

Rule based systems

A rule-based approach is a semi-automatic system which relies on a set of extraction rules which may be provided by the user or generated automatically by the software. It is based on the concept that words, phrases and other linguistic annotations can co-occur in similar linguistic contexts. These contexts can be defined by pre-processing the text to identify noun, prepositional, adverbial, adjectival and verb phrases and other significant syntactic relations such as, subject-verb and verb-object relations. An example of syntactic relations is given by (i) the pattern verb-object, by which the FB type, which is the object of verbs "ingested" or "swallowed", was identified; or (ii) the identification of proper noun in order to get the gender of the injured child.

Acquisition of linguistic patterns by the software requires the definition of a dictionary that sufficiently covered domain information. To enter this task, the lists of words created for the NER approach were supplied to the software. In particular, a set of concepts (child's age and gender, type of FB) was created and each concept was populated using the list of words previously used in the NER approach.

Thus the IE was performed using VisualText [15] whose professional version is free for academic or non-commercial use only. The software automatically generated rules from sample data.

Support vector machine system

By manually annotating a small number of documents with the information to be extracted, a reasonably accurate IE system can be induced from this labeled corpus to a larger collection of text. In order to perform this task, a SVM information extraction system was tested using T-REX [16] which required the implementation of SVM^{perf} [17].

Basically, the SVM is a supervised machine learning algorithm used for classification. When applying machine learning to IE, a learning algorithm learns a model from a set of articles (the training set) which have been previously annotated. Then the model can be used to classify each word in a new document as belonging to one of the target classes (named entity tags).

Classification is based on boundary classification algorithm which is a binary classifier for detecting the boundaries of the entities to be extracted (one classifier for start word and one classifier for end word) [16].

In our case, a training set of 22 articles were preprocessed using ANNIE which provided the following linguistic features: (a) tokenization, i.e. identification of words of different type (words in uppercase and lowercase) and character punctuations; (b) name entity recognition; (c) Part-Of-Speech tagger; i.e. the identification of nouns, verbs, adjectives; (d) Semantic tagger, which was run to assign nouns and eventually verbs and adjectives to the correct category (gender, age and FB type). The annotation text produced by ANNIE was then passed to T-REX, which produced the model tested on the remaining articles.

Performance measures

Information Extraction systems were evaluated calculating Recall, Precision and the F-measure [18]. Precision is defined as the percentage of the correct extracted informa-tion on the extracted information. Recall is defined as the percentage of the correct extracted information on the whole information in the text. Finally, the F-measure is defined as the weighted harmonic mean of Precision and Recall. Rule based and the SVM systems were trained on a set of 22 articles and tested on the remaining articles. A two-fold cross-validation ensured that the rule based system as well as the SVM system were evaluated on all 44 articles as NER. A predicted annotation was considered true if it strictly matched the human annotated tag.

Results

A set of 44 articles provided between 2003 and September 2006 were analyzed. A workflow of the Information Extraction system was showed in Fig. 1. In Table 1, the list of words used to detect the FBs in the free text was reported.

Judging by the performance measures, accuracy increased in the following order of techniques: SVM, rule based system and NER. NER and the rule based system approach performed essentially identically. The overall scores of the named entity recognition system were 79% for Recall and 85% for Precision (see Table 2). Further analysis of these results (Table 3), showed that performance measures are mainly influenced by the degree of correct extraction of gender and age information more than of the FB characteristics. On the contrary, regarding the type of FB, using the rule based approach, we achieved a precision of 66% and a recall of 73%, resulting in F-measure equal to 70%, which was better than the results regarding age information (details are showed in Table 3). Age information results with classical rule based system report an F-measure equal to 58% (see Table 3). A summary of a basic text analysis based on reading articles, restricted to the first 26 articles, was shown in Table 4.

Conclusions

Newspapers, even if they are a limited source of information by themselves, represent a readily available source of data which can provide details that are not always available from traditional public health datasets.

FB injuries in children represent an example of accidents that are usually reported on newspapers, particularly when they result in a fatality [9, 19]. However, also resolved injuries are often reported by local media. In a set of 44 articles gathered from newspapers, 60% of them accounted for non fatal events, thus turning out to be a source of valuable supplementary information, which can be used to integrate injury details provided by clinical records [9].

From article accounts it is not rare to learn about the circumstances surrounding the accident, such as the presence of other children or the presence/absence of parents' supervision, and the external cause of the accident. Of course, such information, when derived from clippings, lack of any form of validation, and they must be taken with great attention.

The process of reading and organizing media clips in a structured manner when a large corpus of newspaper clippings is available requires the implementation of Information Extraction systems, which allows for structuring free text in order to gather information about pre-specified events [10, 11] for subsequent statistical analyses.

In this paper, three Information Extraction techniques were analyzed: (a) a NER system; (a) a rule based system;

(b) a SVM approach. Performance measures were higher using the NER and the rule based systems. Although SVM has emerged as one of the leading trainable models for many classification task [20], the small number of articles related to choking injuries in children influenced the results. Even if 44 articles represent a very limited sample (but injuries are not always so frequent to allow to gather large samples), this work is mainly aimed at showing the most promising and user friendly Information Extraction methods to retrieve Natural Language in a structured form.

Confirming previous results, our work showed auto-matic methods such as SVM are less reliable but suitable for only large volumes of articles [21], whereas rule based and NER systems perform well also working on a small corpus of documents [22] showing that a semi-automatic approach for newspaper clippings is a reasonable choice when dealing with a moderate number of documents.

On the other hand, the rule based system is not very effective in detecting the age of the injured child. This is probably due to the fact that very often age is reported as a number making it difficult to recognize patterns for the detection of such information. Indeed, rules were generated though the learning of syntactic relations such as, subject-verb and verb-object. Thus, following a structured scheme when writing about injuries could be crucial for improving the acquisition of linguistic patterns.

Despite their intrinsic appeal, several limitations are still present in the widespread usage of newspaper clippings. A major problem which arises is that many articles deal with the same injury as well as, on the contrary, that one article can summarize more than one accident. Furthermore, in our case we experienced that more difficult is the task of identifying the activity the child was involved immediately before the accident, which it is often described using circumlocutions.

As suggested by Rahman [23], to be an effective surveillance tool, newspaper articles should be used in combination to traditional clinical data to compensate for inadequacies in the clinical sources. At this regard, establishing relationships with local journalists could aid public health professionals in promoting a public health framework in newspaper coverage. At conclusion, the appeal in terms of timeliness and costefficacy of getting structured information in automatic environment is evident and research should be fostered in the direction of overcoming actual pitfalls in the methodology.

References

1. Centers for Disease Control and Prevention, Updated guidelines for evaluating public health surveillance systems: recommenda-tions from the guidelines working group, in MMWR Recomm Rep. 2001. p. 1–51.
2. Voight, B., et al., Injury reporting in Connecticut newspapers. Inj. Prev. 4(4):292–294, 1998.
3. Horan, J. M., and Mallonee, S., Injury surveillance. Epidemiol. Rev. 25:24–42, 2003.
4. Baullinger, J., et al., Use of Washington State newspapers for submersion injury surveillance. Inj. Prev. 7(4):339–342, 2001.
5. Guard, A., and Gallagher, S. S., Heat related deaths to young children in parked cars: an analysis of 171 fatalities in the United States, 1995–2002. Inj. Prev. 11(1):33–37, 2005.
6. Frost, K., Frank, E., and Maibach, E., Relative risk in the news media: a quantification of misrepresentation. Am. J. Public Health 87(5):842–845, 1997.
7. Chapman, S., and Lupton, D., The fight for public health: principles and practice of media advocacy. London: BMJ. xv, 270, 1994.
8. Fine, P. R., et al., Are newspapers a viable source for intentional injury surveillance data? South Med. J. 91(3):234– 242, 1998.
9. Rainey, D. Y., and Runyan, C. W., Newspapers: a source for injury surveillance? Am. J. Public Health 82(5):745–746, 1992.
10. Zhou, G., et al., Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20(7):1178–1190, 2004.
11. Corney, D. P., et al., BioRAT: extracting biological information from full-length papers. Bioinformatics 20(17):3206–3213, 2004.
12. Zigon, G., et al., Child mortality due to suffocation in Europe (1980–1995): a review of official data. Acta Otorhinolaryngol. Ital. 26(3):154–161, 2006.
13. Saggion, H., et al., Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project. Data Knowledge Eng. 48(2):247–264, 2004.
14. Cunningham, H., et al., GATE: a framework and graphical development environment for robust NLP tools and applications. In 40th Anniversary Meeting of the Association for Computational Linguistics (ACL’02). 2002.
15. Text Analysis International Inc., Integrated development environ-ments for natural language processing. 2001.
16. Iria, J., Ireson, N., and Ciravegna, F.. An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM. In Workshop on Adaptive Text Extraction and Mining 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.
17. Joachims, T., Training Linear SVMs in Linear Time. in, Proceed-ings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). 2006.
18. Makhoul, J., et al., Performance measures for information extraction. In Proceedings of DARPA Broadcast News Workshop, (Herndon, VA), 1999.

19. Ghaffar, A., Hyder, A. A., and Bishai, D., Newspaper reports as a source for injury data in developing countries. Health Policy Plan 16(3):322–325, 2001.
20. Collier, N., and Takeuchi, K., Comparison of character-level and part of speech features for name recognition in biomedical texts. J. Biomed. Inform. 37(6):423–435, 2004.
21. Ananiadou, S., Kell, D. B., and Tsujii, J. I., Text mining and its potential applications in systems biology. Trends Biotechnol, 2006.
22. Marshall, R. J., Comparison of misclassification rates of search partition analysis and other classification methods. Stat. Med. 25 (22):3787–3797, 2005.
23. Rahman, F., Andersson, R., and Svanstrom, L., Potential of using existing injury information for injury surveillance at the local level in developing countries: experiences from Bangladesh. Public Health 114:133–136, 2000.

Fig. 1 Workflow of the information extraction system

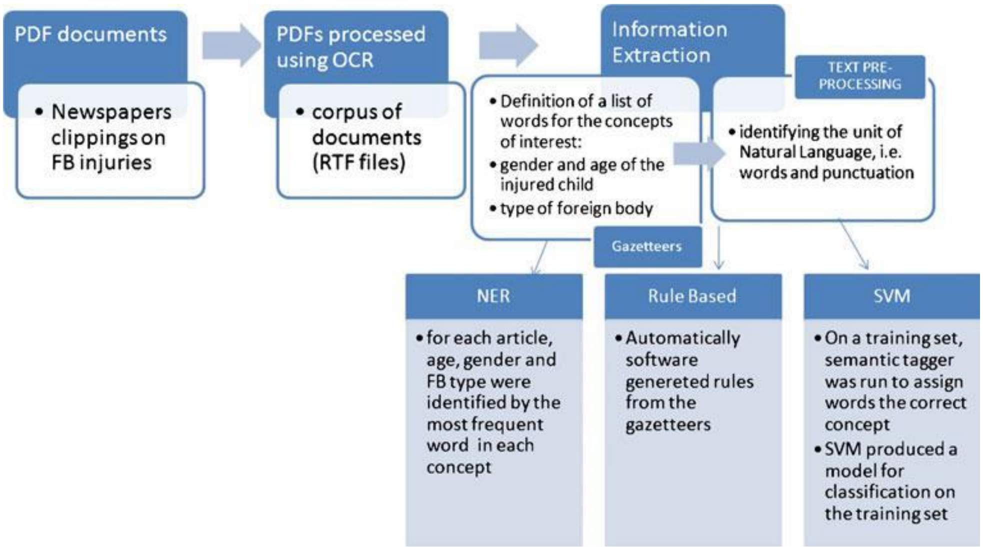
PDF documents
• Newspapers clippings on FB injuries
PDFs processed using OCR
• corpus of documents (RTF files)
Information Extraction
• Definition of a list of words for the concepts of interest:
• gender and age of the injured child
• type of foreign body
TEXT PRE-PROCESSING
• identifying the unit of Natural Language, i.e. words and punctuation
Gazetteers
NER
• for each article, age, gender and FB type were identified by the most frequent word in each concept
Rule Based
• Automatically software genereted rules from the gazetteers
SVM
• On a training set, semantic tagger was run to assign words the correct concept
• SVM produced a model for classification on the training set

Table 1 Partial list of words used for detecting the foreign body which caused the accident

Foreign body type	
Italian word	English translation
Pezzo di mandarino	mandarine orange
Pomodoro	tomato
Pallina	ball
Pezzo di mela	piece of apple
Piece of cheese	mozzarella
fagiolo	bean
Brioche	brioche
Tappo	cap
Stuzzicadenti	wood steack
Pallloncino	balloon
Moneta	coin
Batteria	battery
Sigaretta	piece of cigarette
Patatina fritta	crisp
Pizza	piece of pizza
Carota	carrot
Sassolino	pebble
Boccone	bite
Mais	mais
Pezzo di cibo	food
Nocciolina	nut
Plastilina	plastiline

Table 2 Precision, recall and F-measure performances calculated for each IE technique implemented (NER, rule based system, Support Vector Machine)

IE technique	Performance measures		
	Precision	Recall	F-measure
NER	85%	79%	82%
Rule based system	80%	80%	80%
Support Vector Machine	33%	61%	43%

Table 3 Precision, recall and F-measure performances obtained according to the NER system and the rule based system in detecting gender, age and foreign body information

Information retrieved	IE technique	Performance measures		
		Precision	Recall	F-measure
Gender	NER	90%	82%	86%
	Rule based system	89%	99%	94%
Age	NER	72%	85%	78%
	Rule based system	79%	46%	58%
Foreign body	NER	77%	70%	73%
	Rule based system	66%	73%	70%

Table 4 Summary of a basic text analysis restricted to the first 26 articles

Newspaper title	Article title	Gender	Age	Foreign body type	Outcome	Place	Activity
Il Giornale	Chokes with a mandarin: baby saved by her father	f	1 yrs 2 mon	mandarine orange	alive	Ortonovo (SP)	eating
La Nazione	Dies in the grandparents' arms	m	2 yrs	tomato	deceased	Sesto Fiorentino (FI)	playing
La Gazzetta del Mezzogiorno	Ingestes a small ball. Saved baby of 2 years in Bisceglie	m	2 yrs	ball	alive	Bisceglie (BA)	playing
Il Mattino di Padova	A baby risked to die for choking	m	1 yrs	piece of apple	alive	Fossalta Maggiore (TV)	eating
Città	Choked with a piece of mozzarella	f	4 yrs	mozzarella	deceased	Angri (SA)	eating
Città	Mirko died for a bean	m	1 yrs	bean	deceased	Sapri (SA)	–
Roma	Eats a snack. Baby dies for choking	m	1 yrs 6 mon	brioche	deceased	Banzi (PZ)	eating
La Gazzetta del Mezzogiorno	Baby of 11 months saved	f	11 mon	cup	alive	Leporano (TA)	eating
Il Tirreno	Swallowes a toothpick. Child of 11 years at hospital	f	11 yrs	wood steack	alive	Prato (PT)	eating
Gazzetta del Lunedì	3 years old baby soffocates while blowing up a balloon	f	3 yrs	balloon	deceased	Orvieto (TR)	playing
Il Giorno	Swallows a coin, risks soffocation	f	2 yrs	coin	alive	Vigevano (PV)	–
Provinpavese	Baby swallows a battery	f	10 mon	battery	alive	Roma (RM)	–
Il Mattino	Two years old, soffocated by the ball of a bar billiards	m	2 yrs	small ball	deceased	Brusciano (NA)	playing
Il Girono Varese	Baby swallows a cigarette end. Saved by a stomach pumping	m	2 yrs	piece of cigarette	alive	Lecco (LC)	–
Corriere Veneto	Swallows a crisp, risks to die	m	5 yrs	crisp	alive	Castelfranco (TV)	eating
La Stampa	2 years old baby risks soffocation	m	2 yrs	piece of pizza	alive	Murisengo (AL)	eating
Corriere dell'Alto Adige	15months old baby swallows a piece of carrot. The doctor saves him on the point of death, now he is in a coma	m	1 yrs 3 mon	carrot	alive	Bolzano (BZ)	not clarified
Libero	At 3years swallwos a one Euro coin. T he endoscope lacks, he is risking the life	m	3 yrs	coin	alive	Frosinone (FR)	–
Il Gazzettino	A pebble entered in the nose. Saved on the point of death	m	4 yrs	pebble	alive	Cavallino-Trepori (VE)	playing

–