Automatic Knowledge Extraction and Data Mining from Echo Reports of Pediatric Heart Disease: Application on Clinical Decision Support

Yahui Shi[1], Zuofeng Li[2], Zheng Jia[3], Binyang Hu[1], Meizhi Ju[3], Xiaoyan Zhang[1,*] and Haomin Li[4, 5,*]

[1]School of Life Science and Technology, Tongji University, Shanghai, China
[2]HealthCare, Philips Research China, Shanghai, China
[3]College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhejiang, China
[4]The Children's Hospital of Zhejiang University, Zhejiang, China
[5]The Institute of Translational Medicine, Zhejiang University, Zhejiang, China
[*]Correspondence: hmli@zju.edu.cn; xyzhang@tongji.edu.cn

Abstract. Echocardiography (Echo) reports of the patients with pediatric heart disease contain many disease related information, which provide great support to physicians for clinical decision. Such as treatment customization based on the risk level of the specific patient. With the help of natural language processing (NLP), information can be automatically extracted from free-text reports. Those structured data is much easier to analyze with the existing data mining approaches. In this study, we extract the entity/anatomic site-feature-value (EFV) triples in the Echo reports and predict the risk level on this basis. The prediction accuracy of machine learning and rule-based method are compared based on a manual prepared ideal data, to explore the application of automatic knowledge extraction on clinical decision support.

Keywords: echo reports; natural language processing; knowledge extraction; machine learning; clinical decision support

1 Introduction

Echocardiography (Echo) is one of the most widely used diagnostic tests in cardiology[1]. After examination, an Echo report recording the findings and conclusions is generated by the physician, which is regarded as an important evidence to support the clinical practice. Most electrocardiographic left ventricular hypertrophy (ECG-LVH) studies have used echocardiographic left ventricular mass (Echo-LVM) as the gold standard for evaluating ECG-LVH criteria[2]. Focused cardiopulmonary ultrasonography disclosed unexpected pathology in patients undergoing urgent surgical procedures[3]. When the patient is discharged, those reports are deposited in the database. As time goes on, there are more and more Echo reports being accumulated. In this situation, natural language processing (NLP) technique has been popularly used in

medical domain to facilitate the conversion from free-text clinical records to structured data for analysis/mining[4]. Many studies have focused on such data sources to discover interesting pattern, to find a way to transform current clinical workflow and to improve the clinical quality[5,6,7]. Moreover, integrating NLP tool into decision support system also makes it possible to alert the physician to ambiguities and omissions when a report is generated[8].

We adopt a hybrid approach to extract and organize the anatomic site-related description in the Echo reports. Mining on these converted data is expected to provide evidence support and novel knowledge for clinical practice. In this study, a risk level is predicted for the patient based on the Echo report processed by the NLP module. Machine learning and rule-based methods are compared for their prediction performance.

2 Data and Methods

2.1 Data collection

Currently, more and more patients would like to put their data online for consultant. Some patients voluntarily post their reports to ask the physician for medical advice. In some cases, the consultant information is freely available online (www.haodf.com, http://dxy.com/faq, etc.). At the same time, many hospitals have opened forums for the communication between physicians and patients such as Fuwai hospital, Shanghai Children's Medical Center (SCMC) and Fudan tumor hospital. Some physicians are actively involved in the online consultant. In this study, we use the data from SCMC. For each post, the physician ranks the risk level based on the Echo report contents and the patient's condition. This evaluation ranges from level one to level five as the risk increasing. (For detailed guideline refer to http://www.ibabyheart.com/hazard.html). In this study, those Echo report contents and the risk evaluation results are collected and analyzed.

A home-made python script is used to collect the posts from the Neonatal Congenital Heart Disease Forum, which filters the web pages with the keywords of "color Doppler ultrasound reports" ("彩超报告"). As labeled by xml tags, Echo report contents and the corresponding risk level evaluations are automatically recognized and extracted from the target web pages. The contents are organized in a flat file containing several sections (as shown in Fig. 1).

7062 posts posted before 2015 March were collected from the forum. 3464 posts among them contain both the Echo report contents and risk level evaluation, on which we try to explore our knowledge extraction and data mining approaches.

```
>title

>findings
二、检查描述：

内脏心房正位，心室右袢。右心扩大，主肺动脉无扩张。房间隔近十字交叉处连续性中断，最大缺损径0.9cm，左向右分流。室间隔可见膜部瘤，基底部宽5mm，未见左向右分流，未见其他心内分流。左室收缩功能正常，未见心包积液。二尖瓣前叶裂，宽度约占瓣环的1/3，收缩期见少量返流，余瓣膜结构正常，血流速度正常范围，三尖瓣收缩期见少量返流。 多普勒估测肺动脉收缩压34mmHg。胸骨上窝探查未见异常。
>conclusions
二、超声提示：
1、先天性心脏病
2、原发孔房缺，二尖瓣前叶裂（部分性心内膜垫缺损）
3、室间隔膜部瘤
4、二尖瓣及三尖瓣少量反流
5、右心扩大
>measurements

>others
下面是我女儿的基本信息：
性别：女　　年龄：2岁　　身高：92cm　　体重：13kg　　体表面积：0.57m2
彩超报告：检查结果
>RankLevel
04
#http://www.ibabyheart.com/thread-11004-1-129.html
```

Fig. 1. An example showing file format. Each section is labeled by a title initiating with a right angle bracket (>).

2.2 Free text to structured data

Information extraction and normalization.
A hybrid approach was adopted to extract all data from free text of Chinese Echo reports. The output of each step is the input of the next one. Firstly, we use the software CRF++ (version 0.58) to train a CRF model, which labels the free text reports with three classes: entity (anatomic site), feature and value. Secondly, a series of rules are applied to build the semantic relationships among the text spans. Then the pathological description about the anatomic sites in the raw text is converted to several Entity-Feature-Value (EFV) entries. Each entry consists of semantically related entity, feature and value. Thirdly, the labeled text spans are normalized and coded with the pre-defined ontology for entity and the dictionaries for each class. A dynamic programming-based algorithm is designed to implement the dictionary-look-up approach. The context information is used for the disambiguation of general words. After all the concepts are normalized, each EFV triple can be represented by a set of codes. Those codes are alphabetically sorted and jointed to form a string, which is called EFVCode. Lastly, the EFVCodes that we're interested in are collected to form a standard set of attributes for further analysis. Those attributes can be either Boolean or quantitative (Fig. 2). In this way, each report can be represented by a vector containing values of each attributes. The whole set of reports can be converted into a matrix for data mining, where every record represents a report and every column represents an attribute.

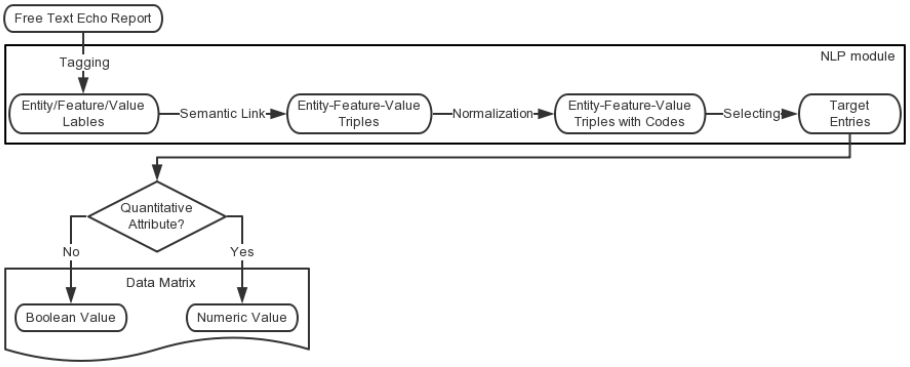

Fig. 2. Flow chart to show the NLP module (upper box) for knowledge extraction and data matrix for further analysis/mining.

IdealData.
For the sake of performance evaluation, 50 Echo reports (10 for each risk level) are randomly selected for manual annotation, generating the ground truth, which is called "IdealData". The details about annotation are described in section 2.3.

Since it can be regarded as the perfect result of knowledge extraction, the IdealData can be used to compare the accuracy of machine learning and rule-based risk prediction. We feed it to the knowledge module to explore its application in clinical decision support.

2.3 Manual annotation

Knowtator plug-in for Protégé is used as the annotation tool for tagging and semantic link. For normalization procedure and target entry selection, manual annotations are organized in flat files. In a former work, two annotators have double annotated 420 templates of adult echo report for training the CRF model, which extracts EFV entries from free-text echo reports. After several rounds of discussion on the difficult cases and differences between two annotators' decisions, the final Inter-Annotator Agreement (IAA) has achieved 95.96% on average. The agreed annotation strategies were taken as the annotation guidelines.

In this study, as time and resources are limited, 5 trained annotators (including the two annotators mentioned before) prepared the IdealData collaboratively. Then the annotations were reviewed and refined by the most experienced annotator, to be consistent with the annotation guidelines.

2.4 Knowledge module

Rule-based model.
Based on the detailed guideline about risk level evaluation, we have built up a rule-based module as the baseline system. This module takes the data matrix (described in

section 2.2) as input and output the risk level evaluation. To build the module, the keywords in the definition of each risk level are picked out to be the criteria, like "complete transposition of great arteries, without pulmonary stenosis" for level five, "double outlet right ventricle" for level four, "atrial septal defect" for level three, "ventricular septal defect" for level two and "diameter of oval foramen less than 4 mm, and age not more than 3 months" for level one, and so on. For each patient, we firstly decide whether he/she belongs to the highest risk level. If the document is classified into this level, the work is done. If not, the patient will be evaluated with the criteria for the next level. If no risk level has been classified, the document will be labeled as 'unclassified' (as shown in Fig. 3.).

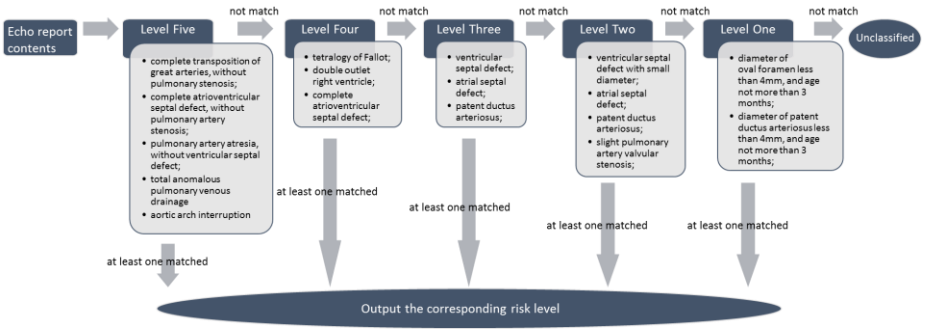

Fig. 3. Workflow to show the rule-based module for risk level evaluation.

Machine learning model.

After machine learning, we get an EFV-based classifier, which takes the extraction results from NLP module as the input (data matrix introduced in section 2.2) to predict the risk level. Fig. 4. shows the workflow of building the machine learning model.

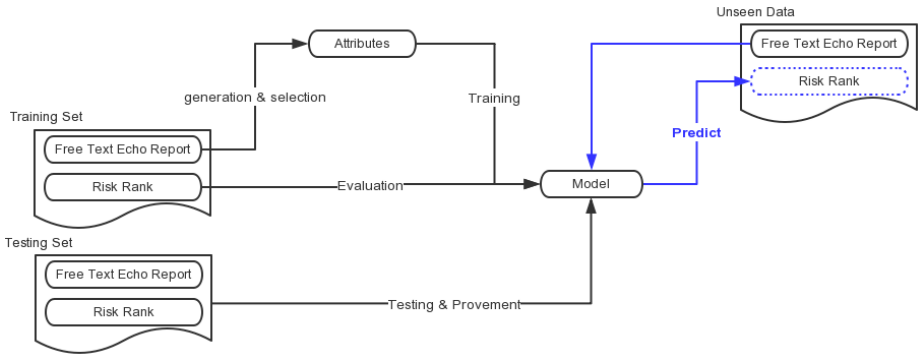

Fig. 4. The workflow of building the machine learning model for risk level prediction.

We use the software Weka (version 3.6) as the tool to train a decision tree for prediction. The algorithm returns readable rules, making it possible for us to compare those rules to the physician's definition, conduct further improvement on both modules, and find some interesting results from the model.

2.5 Evaluation

For the performance of NLP part, the system output is evaluated against the IdealData. The risk level that physician provided in the online replies are used as ground truth, which are also extracted from the web pages during the data collection. By comparing the model prediction with the ground truth, we can calculate the accuracy of the prediction. The rule-based prediction is directly compared with the ground truth to get the performance. While the classifier trained by machine learning is validated through a 10-fold cross-validation approach.

3 Results and discussion

The NLP module to extract the target attributes yields an accuracy of 0.44 in F1-measure (0.51 in precision; 0.43 in recall). Most of the errors are caused by the semantic patterns and clinical terms that are unseen in the training data, which is mainly from the adult Echo reports. This indicates that further optimization is needed for the adaptation of the NLP tool to a new domain.

To explore whether the machine learning method can be used to automatic knowledge extraction, we trained a decision tree based on the IdealData (Fig. 5), whose prediction performance is provided in Table. 1.

```
J48 pruned tree
------------------

三尖瓣收缩期反流 = 未提及
|   房水平分流 = 向右: 四级 (3.0/1.0)
|   房水平分流 = 未提及
|   |   房室瓣开放 = 未提及
|   |   |   二尖瓣位机械瓣功能正常 = 未提及
|   |   |   |   心包胸腔积液 = 未提及
|   |   |   |   |   室间隔回声 = 未提及
|   |   |   |   |   |   室间隔膜周部回声中断 = 未提及
|   |   |   |   |   |   |   肺动脉瓣血流速度 = 未提及
|   |   |   |   |   |   |   |   室间隔延续 = 完整: 二级 (2.0)
|   |   |   |   |   |   |   |   室间隔延续 = 未提及: 五级 (13.0/5.0)
|   |   |   |   |   |   |   |   室间隔延续 = 无中断: 二级 (1.0)
|   |   |   |   |   |   |   |   室间隔延续 = 中断: 五级 (0.0)
|   |   |   |   |   |   |   肺动脉瓣血流速度 = 未见增快: 五级 (0.0)
|   |   |   |   |   |   |   肺动脉瓣血流速度 = 正常: 五级 (0.0)
|   |   |   |   |   |   |   肺动脉瓣血流速度 = 增快: 二级 (2.0)
|   |   |   |   |   |   室间隔膜周部回声中断 = 中断: 三级 (2.0)
|   |   |   |   |   室间隔回声 = 完整: 三级 (1.0)
|   |   |   |   |   室间隔回声 = 中断: 三级 (4.0)
|   |   |   |   |   室间隔回声 = 无明显缺失: 二级 (5.0/1.0)
|   |   |   |   心包胸腔积液 = TRUE: 三级 (0.0)
|   |   |   |   心包胸腔积液 = 未见: 四级 (3.0/1.0)
|   |   |   二尖瓣位机械瓣功能正常 = 正常: 一级 (2.0)
|   |   房室瓣开放 = 可: 一级 (4.0/1.0)
|   |   房室瓣开放 = 正常: 四级 (1.0)
|   房水平分流 = 未见: 一级 (1.0)
|   房水平分流 = 双向: 四级 (2.0)
|   房水平分流 = 双向|向右: 四级 (1.0)
三尖瓣收缩期反流 = 少量: 四级 (2.0)
三尖瓣收缩期反流 = 可见: 五级 (1.0)
```

Fig. 5. The decision tree generated from IdealData using J48 algorithm provided by Weka. The values in the parentheses represent the number of training instances reaching the leaf and that of the misclassified (if any) respectively.

From the decision tree above, we can find it covers some criteria mentioned in the risk level evaluation guideline while referring to some other information. On one hand, this result indicates that the machine learning method is able to find the key information for clinical decision support. On the other hand, with further investiga-

tion, the newly found attributes may provide some novel knowledge for clinical practice.

Table 1. Accuracy of machine learning system

Class	Precision	Recall	F-Measure	ROC Area
Level One	0.7	0.7	0.7	0.794
Level Two	0.429	0.3	0.353	0.519
Level Three	0.714	0.5	0.588	0.716
Level Four	0.286	0.2	0.235	0.575
Level Five	0.211	0.4	0.276	0.439
Weighted Avg.	0.468	0.42	0.43	0.609

The same data is also fed into the rule-based system. However, some definitions about the risk level are still vague, like "slightly defect" and "small diameter". Moreover, as there are many variations in the expression pattern, many criteria for risk level evaluation mentioned by the guideline can't be exactly found in the real-life reports. Thus 44 records in the IdealData can't be classified with the rule-based system. Among the others, 2 patients are correctly assigned to level three, 1 patient is correctly assigned to level two and 3 patients belonging to level four are misclassified as level three.

4 Conclusions and future work

Machine learning is a promising method for clinical decision support compared with the rule-based approach especially when there is no completed knowledge ready. The result indicates that our approach is powerful to facilitate the data mining on clinical free text. The key features and the rules extracted by our method are reasonable and conducive for clinical decision supporting.

For the next step, we will further improve the knowledge extraction and risk prediction accuracy and make the both modules more generalizable and extendable. In the future, the NLP module can be integrated into the report generating system, which alerts the physician when any critical information is omitted in the Echo reports. On the data matrix, other data mining approaches will be explored to utilize its clinical application.

References

1. Maleki, M. & Esmaeilzadeh, M. The evolutionary development of echocardiography. *Iran. J. Med. Sci.* **37,** 222–232 (2012).
2. Rautaharju, P. M. & Soliman, E. Z. Electrocardiographic left ventricular hypertrophy and the risk of adverse cardiovascular events: A critical appraisal. *J. Electrocardiol.* **47,** 649–654 (2014).

3. Botker, M. T., Vang, M. L., Grofte, T., Sloth, E. & Frederiksen, C. A. Routine preoperative focused ultrasonography by anesthesiologists in patients undergoing urgent surgical procedures. *Acta Anaesthesiol. Scand.* **58,** 807–814 (2014).
4. Hughes, K. *et al.* The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J. Pathol. Inform.* **3,** 23 (2012).
5. Krysiak-Baltyn, K. *et al.* Compass: a hybrid method for clinical and biobank data mining. *J. Biomed. Inform.* **47,** 160–170 (2014).
6. Reiner, B. Uncovering and improving upon the inherent deficiencies of radiology reporting through data mining. *J. Digit. Imaging* **23,** 109–118 (2010).
7. Mani, S. *et al.* Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J. Am. Med. Inform. Assoc.* **21,** 326–336 (2014).
8. Bozkurt, S. & Rubin, D. Automated detection of ambiguity in BI-RADS assessment categories in mammography reports. *Stud. Health Technol. Inform.* **197,** 35–39 (2014).