# Automated Reconciliation of Radiology Reports and Discharge Summaries

**Bevan Koopman, PhD[1,2], Guido Zuccon, PhD[2], Amol Wagholikar, PhD[1], Kevin Chu, MBBS FACEM[3], John O'Dwyer[1], Anthony Nguyen, PhD[1], Gerben Keijzers, MBBS PhD FACEM[4]**

**[1]Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia;**
**[2]Queensland University of Technology, Brisbane, QLD, Australia;**
**[3]Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia;**
**[4]Gold Coast Hospital, Gold Coast, QLD, Australia.**

## ABSTRACT

We study machine learning techniques to automatically identify limb abnormalities (including fractures, dislocations and foreign bodies) from radiology reports. For patients presenting to the Emergency Room (ER) with suspected limb abnormalities (e.g., fractures) there is often a multi-day delay before the radiology report is available to ER staff, by which time the patient may have been discharged home with the possibility of undiagnosed fractures. ER staff, currently, have to manually review and reconcile radiology reports with the ER discharge diagnosis; this is a laborious and error-prone manual process. Using radiology reports from three different hospitals, we show that extracting detailed features from the reports to train Support Vector Machines can effectively automate the identification of limb fractures, dislocations and foreign bodies. These can be automatically reconciled with a patient's discharge diagnosis from the ER to identify a number of cases where limb abnormalities went undiagnosed.

## Introduction

The misdiagnosis of a patient's true clinical condition due to misinterpretation of radiological evidence by the treating clinician is an occasional problem in hospital emergency departments. There is always a time delay between reporting of the radiologist and clinical treatment by the Emergency Room (ER) clinician. The large amount of manual processing of unstructured text is one of the main issues that can be resolved by technology enabled solutions.

A good example of a misdiagnosis issue is the identification of subtle limb abnormalities (fractures, dislocation or foreign bodies). Radiological evidence of limb abnormalities, when subtle, can be missed by clinicians working in the ER. The reporting of a abnormalities by a radiologist may not occur in real time and therefore may not be available to the clinician treating a patient. Consequently, patients may be sent home without appropriate treatment and follow up. A study by Cameron[1] reported that 2.1% of all fractures were not identified on their initial presentation to the ER. Furthermore, Sprivulis and Frazer[9] reported that 1.5% of all x-rays have abnormalities not identified in the ER records. Similarly, Mounts et al.[5] reported that 5% and 2% of the x-rays of the hand/fingers and ankle/foot from a paediatric ER had fractures missed by the treating clinician. Although small, these percentages are not insignificant.

The need to reduce errors is well recognised[4,7,8]. To ensure a diagnosis is not missed, radiology reports are commonly checked and patient records are reviewed, but this may not happen until days after the initial presentation. The current clinical practice of identifying limb abnormalities from radiology reports is highly labour intensive and is subject to human error or omissions. There is a need to streamline the process of identifying missed abnormalities for better patient outcomes. Technology enabled solutions that can streamline the diagnosis identification would certainly improve efficiency in the existing process.

Previous work has focused on automatically detecting fractures from free-text radiology reports. De Bruijn et al.[3] considered acute fractures of the wrist and reported that a Support Vector Machine algorithm (SVM) was able to identify fractures in free-text radiology notes, achieving an overall F-measure of 91.3%. While, Thomas et al.[10] developed a text search algorithm that accurately classified radiology reports into the categories "fracture", "normal" and "neither normal nor fracture". Zuccon et al.[15] have studied Naive Bayes and Support Vector Machines based classifiers for the identification of limb abnormalities. They have shown that machine learning techniques coupled with both word and semantic features are very effective for this task, achieving an overall F-measure of 92.3%. Their evaluation however was limited to a sample of 99 radiology reports from a single hospital radiology service.

Table 1: Three different datasets of radiology reports, along with the number of normal and abnormal cases as identified through our annotation process. The average document length for free-text reports in each dataset is also recorded: the large difference in average length between GCH and RBWH/RCH may be due to differences in reporting language and style conventions.

| Dataset | Description | #Reports | Normal | Abnormal | Avg. Doc. Len. |
|---|---|---|---|---|---|
| RBWH | Royal Brisbane & Womens' Hospital (adult) | 1,480 | 58% | 42% | 52 words |
| RCH | Royal (Brisbane) Childrens' Hospital (child) | 498 | 66% | 34% | 50 words |
| GCH | Gold Coast Hospital (adult child 38%) | 400 | 62% | 38% | 27 words |

In this paper, we build upon the work of Zuccon et al.[15] and we experiment with the automatic classification of free-text radiology reports for identifying abnormalities of limb structures using machine learning algorithms and features such as bigrams formed by stemmed tokens, negations, and SNOMED-CT concepts extracted from the free-text. While previous work has shown promise, it does not address a number of important areas for the practical use of such techniques in a clinical setting. We outline these below and highlight the contribution this study makes in addressing each.

1. Where the majority of the existing methods only used term-based features we extract medical concepts from the SNOMED-CT ontology and exploit these for training the classification model.
2. Previous studies used datasets that were small and homogenous[10,3,15]. Instead, in this study, we use a larger set of reports taken from different hospitals and different ages (adults vs. children). An important requirement for the general applicability of these methods is the robustness of the models across different hospital datasets; i.e., how well would one method, developed using one hospital's set of reports work when deployed at another hospital, where particular institutional conventions may result in different language and authoring styles in radiology reports. In addition, radiology reports for children and adults may also differ. To address this issue, we evaluate our method on three different sets of radiology reports from three different hospitals (adults and children). These types of heterogeneous data sources were not considered in previous studies. In this study, we empirically show, via different training/test combinations, that the method developed on one hospital's reports (with differing conventions and patient cohorts) can be applied to another with marginal loss in effectiveness.
3. Finally, and perhaps most importantly, we investigate how the classification of radiology reports can be used in a real clinical setting to reconcile the radiology diagnosis with that of the patient's discharge diagnosis from the ER, thereby identifying a number of cases where limb abnormalities may have been undiagnosed. Thus we study an end-to-end application of natural language processing and machine learning to aid clinicians in the identification of undiagnosed limb abnormalities.

## Materials and Methods

### Data

A set of 2,378 free-text radiology reports of limb structures was acquired from the Emergency Department of three large Australian public hospitals (adult, children and mixed adult/children). Ethics approval was granted by the Human Research Ethics Committee at Queensland Health to use the non-identifying data. Free-text reports were short in length, containing on average 47 words, and an (unstemmed) vocabulary comprising 4846 unique words. Details of the three datasets are outlined in Table 1. Free-text reports in the GCH dataset were found to be on average consistently shorter than those in the other two datasets: this may indicate differences in reporting style and conventions between the hospital sites.

Free-text reports were manually annotated by an Emergency Medicine Registrar and a Medical Officer as being either:

**"normal"** — the radiography does not exhibit a fracture, dislocation or presence of a foreign body; or

**"abnormal"** — some fracture, dislocation or foreign body was found.

A software tool was developed to assists clinicians in the recording of their interpretation and to highlight the portion of text in the report that lead to their interpretation.

Initially, assessors agreed on the annotations of 2,215 out of the 2,378 reports. A senior Staff Specialist in Emergency Medicine was then asked to act as third assessor and resolve disagreements. The distribution of normal and abnormal cases across the three datasets is reported in Table 1. The Fleiss' kappa ($\kappa$) calculated on the initial set of annotations provided by the two first assessor was 0.85, thus exhibiting strong inter-rater reliability.

### Automatic Feature Extraction and Weighting

Machine learning algorithms require documents to be described by features. The text analysis capabilities of the Medtex tool were developed to automatically extract features from the free-text radiology reports[6]. Medtex is a text analysis system that has been previously used for classify cancer-notifiable pathology reports and produce a minimum set of synoptic factors. A wide range of features were initially extracted, including:

- token, i.e., a word found in a report;
- punctuation;
- token stem, i.e., the stemmed version of a word contained in a report;
- token negation, i.e., if a token or phrase was explicitly negated (e.g., "no fracture"); the Medtex implementation of the ConText algorithm[2] was used to identify negations in free-text;
- token stem bi-gram, i.e., a pair of adjacent stemmed words as found in a report;
- token stem tri-gram, i.e., a 3-tuple of adjacent stemmed words contained in a report;
- SNOMED-CT concepts extracted from the text of the report;
- the fully specified terms of extracted SNOMED-CT concepts restricted to morphologic abnormalities and disorders;
- SNOMED-CT concept bi-gram, i.e., a pair of adjacent SNOMED-CT concepts as found in a report.

While a number of these features are commonly used for the classification of free-text documents, the use of SNOMED-CT features have not been widely evaluated by previous works on classification of radiology reports. To our knowledge, only Zuccon et al. [15] investigated these features but their evaluation was limited to a small sample of 99 radiology reports. In this work, SNOMED-CT concepts were extracted by annotating the radiology reports with the MetaMap: a natural language processing tool that identifies medical concepts mentions in free-text. The actual feature used in training the classifier was the SNOMED-CT concept ids found my MetaMap. Previous empirical results have shown that SNOMED-CT concepts, in particular those referring to abnormalities (e.g., fracture, dislocation, etc.) and disorders (e.g., fracture of bone, traumatic injury, etc.), provide valuable evidence for representing free-text radiology report data[15]. Table 2 provides an example of feature sets extracted from the free-text of the radiology reports.

Table 2: Features extracted from two example free-text radiology reports; a 1 corresponds to the feature being present.

| | **Features** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | stem | | | | | | stemBigram | | | | | concept | | | ... | |
| | moder | soft | .. | swell | dorsal | disloc | moder_soft | soft_tissue | .. | tissue_swell | disloc_present | 298349001 | .. | 108367008 | ... | Abnormal? |
| Report 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | ... | 0 |
| Report 2 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | ... | 1 |

### Automatic Classification and Evaluation Methodology

To classify radiology reports we used the Weka toolkit API[14] and the corresponding implementation of the Sequential Minimization Optimization (SMO) classifier. The SMO classifier is a support vector machine (SVM) algorithm where training is performed according to the sequential minimal optimisation algorithm and a polynomial kernel is used. The parameters of all classifiers were set to the default values (see Witten et al.[14] for details).

To explore the effectiveness of the machine learning classifier for identifying limb abnormalities we conducted three sets of experiments.

*Experiment 1.* We combine all three datasets of Table 1 and use the 10-fold cross validation methodology to evaluate the classification algorithms. In this methodology, the dataset is randomly divided into 10 stratified folds of equal dimension (in our case nine folds will contain 238 reports, while the remaining fold will contain only 236 reports). The model for each classifier was then learnt on nine of these folds, leaving one fold out for testing the model. The process was repeated by selecting a new fold for testing, while a new model was learnt from the remaining folds. Classification performances were then averaged across the folds left out in each iteration. The aim of this experiment was to evaluate the effectiveness of a classifier learnt on the whole combination of datasets.

*Experiment 2.* We consider the reports from each hospital dataset separately. For each dataset, we use the 10-fold cross validation methodology to evaluate the classification algorithms specific to that dataset. Thus, experiments on the larger RBWH dataset are characterised by larger folds than those on the remaining two smaller datasets. The aim of this experiment was to evaluate the effectiveness of classifiers specifically learnt on individual datasets and thus individuate whether a dataset is more challenging than others for automatic classification (and what the possible causes for this are).

*Experiment 3.* We performed a split train/test evaluation: train on one hospital's reports and test on another hospital's reports. This procedure was repeated for all combinations of hospitals and included training on reports from two hospitals and testing on those from another. The aim of this last experiment was to evaluate the robustness of our method across reports from different hospitals.

As a baseline for comparison against our machine learning method, we included a keyword spotting system, which resembled the method by Thomas et al.[10] and Wagholikar et al.[13]. A set of regular expressions were defined based on common phrases or terms that identify an abnormality; these were based on discussion with a senior Staff Specialist in Emergency Medicine. (Details of the regular expressions are provided in Appendix .)

Two evaluation measures were considered: precision and recall (also called positive predictive value and sensitivity, respectively). Precision is the fraction of positively classified reports that contain abnormalities, while recall is the fraction of actual abnormalities that were positively classified. In addition, to provide a single, overall evaluation measure, precision and recall are combined into a third evaluation measure, F-measure.

### Reconciliation of Radiology Reports with Emergency Room Discharge Diagnosis

Using the methods described here we are able to automatically identify abnormalities from a patient's radiology report. The benefit of such a method is the ability to reconcile the abnormality classification with the discharge diagnosis from the ER to ensure that an abnormality did not go unrecognised and the patient discharged without proper treatment. To demonstrate the utility of this, we reconciled all the radiology reports used in the classification task with the ER discharge diagnosis ICD-10 code. If the ER discharge diagnosis ICD-10 code matched a predefined set of "abnormal" codes then the patient was marked as abnormal; else they were marked "normal".* (The full list of ICD-10 codes considered as abnormal was provided by an ER clinician (KC) and is provided in Appendix .) Patients that had a abnormal radiology classification using our automated method but did not have any abnormality related ICD-10 code recorded in the ER discharge diagnosis were *flagged* as possible misdiagnosis for immediate followup.

*Note that in Australian Emergency Departments ICD-10 codes are used as a diagnostic classification and are not used for billing purposes.

Table 3: Classification results for each of the three datasets, comparing the proposed machine learning method (SVM) against the keyword baseline. The percentage change in F-measure shows the improvement of SVM over the keyword baseline.

| Dataset | Method | Precision | Recall | F-measure |
|---|---|---|---|---|
| RBWH | Keyword | 0.73 | 0.61 | 0.67 |
| | SVM | 0.88 | 0.91 | 0.89 (+33%) |
| RCH | Keyword | 0.74 | 0.78 | 0.76 |
| | SVM | 0.96 | 0.94 | 0.95 (+25%) |
| GCH | Keyword | 0.94 | 0.58 | 0.72 |
| | SVM | 0.96 | 0.94 | 0.95 (+32%) |
| All | Keyword | 0.76 | 0.64 | 0.69 |
| | SVM | 0.92 | 0.92 | 0.92 (+33%) |

Table 4: Split dataset training/testing F-measure results. Grey shaded cells represent cross validation results; all other results are train/test.

| | | Testing | | | |
|---|---|---|---|---|---|
| | | GCH | RBWH | RCH | All |
| Training | GCH | 0.95 | 0.80 | 0.88 | |
| | RBWH | 0.84 | 0.89 | 0.85 | |
| | RCH | 0.87 | 0.81 | 0.95 | |
| | GCH+RBWH | | | 0.88 | |
| | GCH+RCH | | 0.80 | | |
| | RBWH+RCH | 0.84 | | | |
| | All | | | | 0.92 |

**Classification Results and Discussion**

The overall classification results (Experiments 1 and 2) are shown in Table 3. F-measure was used as the overall effectiveness measure and the percentage change in F-measure shows the improvement of the machine learning method over the keyword baseline. For all hospital datasets, the SVM method outperformed the keyword baseline in all evaluation settings. This is consistent with previous results[15]. In addition, the keyword baseline had more variance across datasets (lower on RBWH, higher on RCH), whereas the SVM method was more stable across datasets. For the SVM method, both precision and recall were of similar value, indicating that the errors that did occur were a mixture of false positive and false negatives.

**Split Dataset Training and Testing**

The split training/testing F-measure results (Experiment 3) are shown in Table 4. The first column in the table corresponds to the dataset used in training; the first row corresponds to the dataset the model was tested on; grey shaded cells are the previous cross validation results repeated from Table 3. Comparison with the grey cell cross validation results shows that there was some loss in effectiveness when applying models across datasets (e.g., F-measure of RBWH was 0.89 for cross validation and 0.80–0.81 when trained on other datasets). However, even with some loss in effectiveness, the results indicate that the models were still robust when applied across datasets (certainly compared to the results of the keyword baseline).

Combining datasets for training (e.g., train on GCH+RBWH, test on RCH) showed no significant benefit over using a single dataset to train (e.g., train on RBWH, test on RCH). This result reveals that simply adding more training data

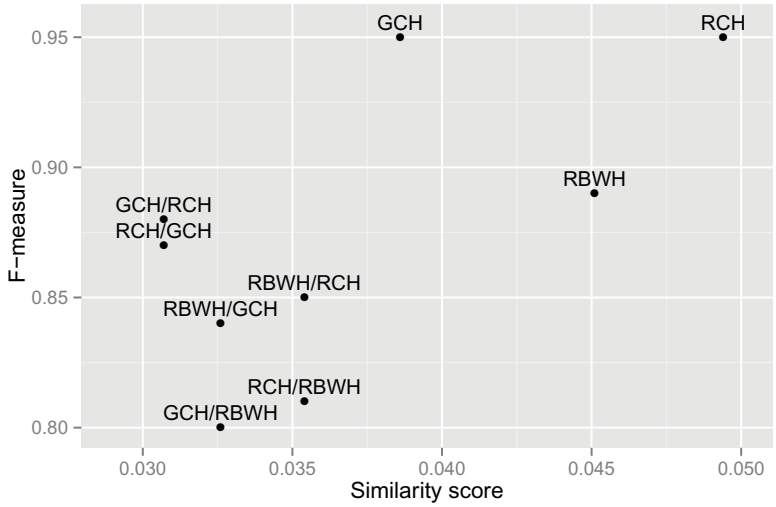



Figure 1: Comparison of similarity between datasets (or itself) and F-measure effectiveness of the classifier. Points with two datasets represents a different train/test combinations RCH/GCH = train on RCH test on GCH), while points with a single dataset (e.g., RCH) are the cross validation results for a single dataset.

does not lead to immediate improvement in effectiveness. Instead, effectiveness was more influenced by the particular dataset used for training. For example, when testing on GCH, it was better to train on RCH than RBWH, even though RBWH was a larger dataset. Based on this finding we set out to understand how similar each of the three datasets were to each other in order to explain the differences in effectiveness in training/testing combinations.

**Dataset Similarity and its Effect on Performance**

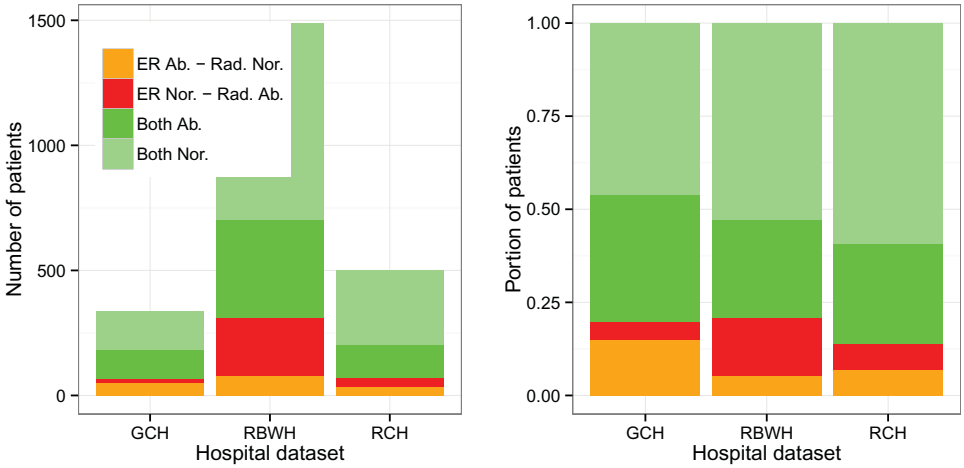
A pairwise similarity calculation was made between the three datasets. This was done by comparing the similarity of every document in one dataset to every other document in another dataset and recording the overall mean similarity. The similarity measure between two individual documents can be calculated by taking the cosine angle between the two documents's term vectors[†] — a standard approach applied in information retrieval when comparing text[11]. Note that the average similarity of a dataset to itself (e.g., GCH vs. GCH) can actually be interpreted as a cohesiveness measure: how similar reports in the dataset are to each other. To understand the similarity results in light of the classifier effectiveness we provide a plot of similarity vs. F-measure in Figure 1 and discuss this in further detail below.

We first consider the cohesiveness of each dataset with itself (i.e., single dataset points). The child-only reports of RCH were the most cohesive, while the mixed adult/child reports of GCH were understandably the least cohesive; adult reports from RBWH were in between. The cross validation training on these datasets obtained the best F-measure — obviously it is best to train and test on same dataset. For these three cross-validated datasets, similarity did not correlate with F-measure (e.g., GCH still had a high F-measure but the lowest similarity score).

Comparing across datasets, any combination containing both RCH (children) and GCH (mixed) had the lowest similarity yet produced the best F-measure (and effectiveness was similar for both RCH/GCH and GCH/RCH). In contrast, for the other four dataset combinations that contained RBWH, the F-measure was lower. In addition, there was a large difference in F-measure for swapping the training/test combination, i.e., RBWH/RCH was lower than RCH/RBWH. For all four combinations that involved RBWH, training on RBWH was better than testing on RBWH. An immediate

[†]A term vector $\vec{d}$ for document $d$ is an $n$ dimensional vector, where $n$ is the size of the (stemmed) vocabulary. Each element in $\vec{d}$ is the TF-IDF weight of a stemmed term from the vocabulary in the document $d$.

Figure 2: Breakdown of different diagnoses combinations for each hospital dataset (left absolute, right normalised). Red (ER Normal & Radiology Abnormal) indicates the patient had a "abnormal" classification from radiology and a "normal" diagnosis from ER. For such cases, the patient was *flagged* as a possible missed abnormality.




explanation for this could be the larger training size of RBWH; however, the aforementioned GCH/RCH combinations showed the best effectiveness with much less training data. Overall, similarity (as measured here) was not correlated strongly with F-measure.

These results highlight that there are differences in terms of similarity and effectiveness for training/testing combinations, with some datasets being easier to classify than others. Variance in training data (as represented by smaller similarity scores for GCH and RCH) provides the model with mix of data for learning different reporting styles; while larger datasets (i.e., RBWH) are better for training but not testing. This combination of variance (i.e., similarity) and size of datasets needs to be considered when selecting appropriate training sets. Importantly, however, the overall differences in terms of effectiveness (F-measure) are small. Thus the classification models are generally robust across these different types of datasets — an important requirement for their use in a real clinical setting.

**Reconciliation Results and Discussion**

The reconciliation process involved checking the classification of a patient's radiology report with their ICD-10 discharge diagnosis from the ER. Four different ER / Radiology combinations were possible: 1) Emergency Abnormal & Radiology Normal; 2) Emergency Normal & Radiology Abnormal; 3) Both Abnormal; 4) Both Normal. For case 2) (where the patient had a "abnormal" classification from radiology and a "normal" diagnosis from ER) the patient was *flagged* as a possible missed diagnosis case. The breakdown of these four combinations for each hospital dataset is shown in Figure 2. The majority of patients had no abnormality recorded in both radiology and ER (light green, Both Normal), followed by patients with abnormalities recorded in both radiology and ER (green, Both Abnormal). A small number of cases were found with no abnormality in radiology but an abnormality in ER (orange, ER Abnormal – Radiology Normal); this occurred when the ER clinician suspected a condition but this turned out to be negative from the radiological assessment (and was, therefore, not an area of major concern). Finally, the number of patients flagged (red), out of the total number of patients, were: GCH 16 / 400 (4%), RCH 26 / 498 (5%) and RBWH 232 / 1480 (16%). The number of flagged cases was considerably higher than previous studies on quantifying missed fracture rates [1] (especially for the RBWH dataset). To understand the reason behind this, we performed a manual analysis of all ER Normal – Radiology Abnormal (red) cases, reviewing both the radiology report, discharge diagnosis ICD-10 code and any associated ER notes. (The judgements were primary provided by the clinical author, KC.)

Our manual analysis showed that the ER discharge diagnosis ICD-10 code was often ambiguous — it may have

Table 5: Categories used in the manual review of all flagged cases (i.e., Emergency Normal & Radiology Abnormal).

| Category | Definition | Example Diagnosis | Comment/Action Required |
|---|---|---|---|
| Unrelated Diagnosis | The patient was assigned an ICD-10 diagnosis unrelated to a possible abnormality. | *Gout*, *Self-harm*, *Congestive heart failure* | Case requiring review by ER clinician. |
| Related Diagnosis | A condition was recorded that could relate to / cover an abnormality but it was not certain. | *Crush injury*, *Laceration*, *Strain or sprain* | Case requiring review by ER clinician. |
| More General Diagnosis | More general condition that could cover an abnormality was coded, often as a result of imprecise ICD-10 coding. These cases should meet the condition "fracture ISA ⟨ICD-10 Diagnosis⟩". | fracture ISA *Injury of the lower foot*. | Coding issue, currently requiring review by ER clinician but should address coding issue in the long term. |
| Missed Diagnosis | Real case of where ED clinician may have missed a limb abnormality. | *No injury found*, *Patient did not wait*. | Cases for actual follow up by ER clinician. |

covered a fracture diagnosis (and thus an abnormality) but this was not explicit. For example, a patient with a fracture noted in their radiology report was discharged with *S57* (Crushing injury of forearm). For this case, the clinician's judgement was that it was not certain whether the crush injury actually indicated a fracture and therefore whether the ER clinician actually knew a fracture was present. *S57* could not be added to the abnormal list as crush injury does not imply a fracture or other limb abnormality as defined here; indeed, there were many patients diagnosed with this code who did not have a fracture finding. Another common case was a discharge diagnoses of "strain or sprain" where an uncertain or very minor fracture was indicated in the radiology reports. For such cases, the ER clinician may have treated the patient (and therefore recorded the diagnosis) as having only a sprain/strain because the fracture was too minor or uncertain. Other flagged cases had a discharge diagnosis completely unrelated to limb injuries; e.g., a ICD-10 code representing *self harm* or *congestive heart failure*. To better understand the different flagged cases, all 274 were manually reviewed and assigned one of four categories described in Table 5. Note that all of the categories represent the situation where a limb abnormality (as defined in this work) *may* have been undiagnosed in ER; however, only the Missed Diagnosis category represents the situation where a limb abnormality was *certainly* missed.

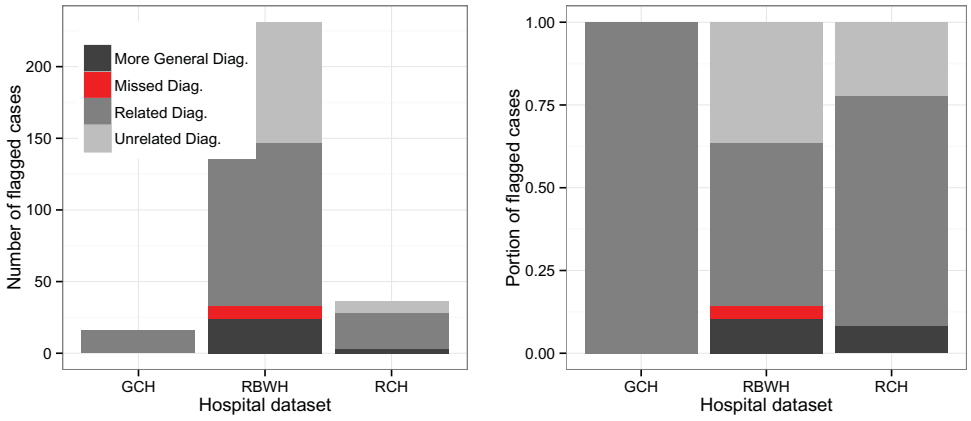
The distribution of these different flagged categories, for each dataset, is shown in Figure 3. The most common category was a Related Diagnosis (e.g., the crush injury example). (All of the GCH and most of the RCH patients fell into this category.) The RBWH dataset contained a larger portion of both Unrelated Diagnosis and More General Diagnosis. This highlights the differences in the way the ICD-10 codes are assigned at different hospitals and how this might affect the use of these codes (our study being but one example of this). A total of 9 genuine Missed Diagnosis cases were identified in the RBWH dataset; these were either *No injury found* or *Patient did not wait*.

Clearly the way ICD-10 codes are assigned affects the reconciliation process, with many flagged cases being the situation where the ER clinician was aware of the abnormality but this fact was not conveyed in the ICD-10 code. However, even given this issue, if the clinician only had to review the 274 flagged cases, this would represent only 11% of the 2,378 reports they previously had to be reviewed — still representing a significant time saving.

## Conclusions

We described a set of techniques to identify limb abnormalities from free-text radiology reports. The empirical evaluation showed that these methods are highly effective and that, importantly, they are robust across hospital datasets that are different in both size and similarity. We further show that the automatic classification can be used to reconcile the radiologist's finding with the ICD-10 discharge diagnosis from the Emergency Room. Using this method, a number of potentially undiagnosed limb abnormalities were identified. A thorough manual analysis of these cases showed that some may be cases where the ICD-10 discharge diagnosis was ambiguous (highlighting the need for accurate ICD-10 coding in ERs); however, some genuine missed diagnoses were uncovered by the automated reconciliation process. Overall, the savings for a clinician were significant with only 11% of the entire dataset now requiring manual review.

Figure 3: Breakdown of different "flagged" cases (radiology abnormal but ER discharge diagnosis normal) according to the four different categories outline in Table 5. (Left plot show the absolute number of cases; the right plot shows the normalised portion of flagged cases.)




As such, the system is part of a pilot study in the Emergency Room of the Royal Brisbane and Women's Hospital.

While the final discharge diagnosis is recorded as an ICD-10 code, the ER clinician may also provide a short description. Although the ICD-10 code may be ambiguous (e.g., crush injury), the clinician's description can contain an explicit mention of an abnormality. This additional source of information, in combination with the ICD-10 code, could be exploited to provide a better classification of the discharge diagnosis. In fact, similar machine learning techniques to those we have described for classifying free-text radiology reports could be adapted to classifying ER notes [12]. The development and evaluation of such a method is an immediate area of future work.

Finally, this study has focused specifically on limb abnormalities described in radiology reports; however, the methods are not specific to this situation. Other types of abnormalities (e.g., presence of cancers) are currently being investigated and the methods are also being applied to the detection and reconciliation of different conditions mentioned in pathology reports (e.g., reconciling the antibiotic given to a patient against the antibiotic sensitivities identified in a microbiology report).

**Acknowledgement**

This research was supported by the Queensland Emergency Medicine Research Foundation Grant, EMPJ-11-158-Chu-Radiology.

**References**

[1] M. Cameron. Missed fractures in the emergency department. *Emerg Med (Fremantle)*, 6:3, 1994.

[2] W. Chapman, D. Chu, and J. Dowling. Context: An algorithm for identifying contextual features from clinical text. In *Proceedings of the Workshop on BioNLP 2007*, pages 81–88. ACL, 2007.

[3] B. De Bruijn, A. Cranney, S. O'Donnell, J. Martin, and A. Forster. Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports. *JAMIA*, 13(6):696–698, 2006.

[4] M. R. James, A. Bracegirdle, and D. W. Yates. X-ray reporting in accident and emergency departments – an area for improvements in efficiency. *Arch Emerg Med*, 8:266–270, 1991.

[5] J. Mounts, J. Clingenpeel, E. M. andE. Byers, and Y. Kireeva. Most frequently missed fractures in the emergency department. *Clin Pediatr (Phila)*, 50:183–186, 2011.

[6] A. Nguyen, J. Moore, M. Lawley, D. Hansen, and S. Colquist. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In *Health Informatics Conference*, pages 117–124, 2011.

[7] M. Saab, J. Stuart, P. Randall, and S. Southworth. X-ray reporting in accident and emergency departments – reducing errors. *Eur J Emerg Med*, 4:213–216, 1997.

[8] E. Siegel, G. Groleau, B. Reiner, and T. Stair. Computerized follow-up of discrepancies in image interpretation between emergency and radiology departments. *J Digit Imaging*, 11:18–20, 1998.

[9] P. Sprivulis and A. Frazer. Same-day x-ray reporting is not needed in well supervised emergency departments. *Emerg Med (Fremantle)*, 13:194–197, 2001.

[10] B. Thomas, H. Ouellette, E. Halpern, and D. Rosenthal. Automated computer-assisted categorization of radiology reports. *American Journal of Roentgenology*, 184(2):687–690, 2005.

[11] C. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[12] M. Vassiliou, A. Jimeno Yepes, M. Gergtz, J. Knott, R. Wynne, and K. Verspoor. Improving consistency of emergency department triage categorisation: Machine learning applied to clinical notes. In *hic2014*, 2014.

[13] A. Wagholikar, G. Zuccon, A. Nguyen, K. Chu, S. Martin, K. Lai, and J. Greenslade. Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *The Australasian medical journal*, 6(5):301, 2013.

[14] I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[15] G. Zuccon, A. S. Wagholikar, A. N. Nguyen, L. Butt, K. Chu, S. Martin, and J. Greenslade. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed-ct ontology. *AMIA Summits on Translational Science Proceedings*, 2013:300, 2013.

### Regex Rules for Keyword Spotting

List of regular expressions used to implement the keyword baseline method. In addition, if a match is found then the matching text is also tested for negation (e.g., "no fracture"), in which case the report is reported as "normal", i.e., no fracture or other abnormality found.

```
"\bfracture", "\bno\b", "\bold\b", "\bfollow[\\s]*up\b", "\bx[\\s]*ray\b", "\bdislocation\b", "\bfb\b",
"\bosteomyelitis\b", "\bosteoly", "\bdisplacement\b", "\bintraarticular extension\b", "\bforeign body\b",
"\barticular effusion\b", "\bavulsion\b", "\bseptic arthritis\b", "\bsubluxation\b", "\bosteotomy\b",
"\bcallus\b", "\bno\b[a-z\\s]+\bfracture"
```

### ICD-10 Abnormal Codes

Set of ICD-10 codes that indicate an abnormal discharge diagnosis from the emergency room.

```
S03.2, S13.10, S03.0, S93.0, S93.30, S73.00, S83.10, S83.0, S93.10, S33.2, S33.10, S23.10, S43.1, S53.10, S63.10,
S53.18, S43.3, S43.2, S63.0, S02.1, S62.1, S42.00, S02.9, S02.4, S02.6, S02.2, S02.3, S02.5, S72.40, S72.00, S72.10,
S72.3, S72.04, S92.0, S92.9, S92.2, S02.0, S42.40, S42.3, S42.20, S62.2, S12.9, S12.8, S82.0, S32.4, S32.5, S32.83,
S52.50, S52.8, S52.4, S62.0, S42.10, S62.5, S82.81, S82.6, S82.5, S82.88, S82.28, S82.82, S82.18, S22.5, S32.00,
S22.3, S22.2, S22.00, S52.20, S82.4, S82.3, S62.6, S92.4, S92.5, M84.49, M86.99, M91.1, M93.0, Z47.8
```