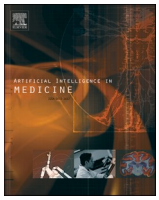
Artificial Intelligence In Medicine 128 (2022) 102281

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



# Radiology report generation for proximal femur fractures using deep classification and language generation models☆

Olivier Paalvast[a,*], Meike Nauta[a,b,**], Marion Koelle[b], Jeroen Geerdink[c], Onno Vijlbrief[c], Johannes H. Hegeman[a,c], Christin Seifert[a,b]

[a] University of Twente, Enschede, the Netherlands
[b] University of Duisburg-Essen, Essen, Germany
[c] Hospital Group Twente, Almelo & Hengelo, the Netherlands

ARTICLE INFO

*Keywords:*
Proximal femur fractures
Radiology report generation
Fracture classification
Radiology language model
User study

ABSTRACT

Proximal femur fractures represent a major health concern, and substantially contribute to the morbidity of elderly. Correct classification and diagnosis of hip fractures has a significant impact on mortality, costs and hospital stay. In this paper, we present a method and empirical validation for automatic subclassification of proximal femur fractures and Dutch radiological report generation that does not rely on manually curated data. The fracture classification model was trained on 11,000 X-ray images obtained from 5000 electronic health records in a general hospital. To generate the Dutch reports, we first trained an embedding model on 20,000 radiological reports of pelvic region fractures, and used its embeddings in the report generation model. We trained the report generation model on the 5000 radiological reports associated with the fracture cases. Our report generation model is on par with state-of-the-art in terms of BLEU and ROUGE scores. This is promising, because in contrast to those earlier works, our approach does not require manual preprocessing of either images or the reports. This boosts the applicability of automatic clinical report generation in practice. A quantitative and qualitative user study among medical students found no significant difference in provenance of real and generated reports. A qualitative, in-depth clinical relevance study with medical domain experts showed that from a human perspective the quality of the generated reports approximates the quality of the original reports and highlights challenges in creating sufficiently detailed and versatile training data for automatic radiology report generation.

## 1. Introduction

Proximal femur fractures represent a major public health concern around the globe. These fractures are predominantly seen in the elderly population, where they are one of the most common contributors to both hospitalisation and mortality [1]. The one-year mortality is conservatively estimated at 20% [2], where the life-time risk of sustaining a hip fracture varies between 4.6–11% and 13.9–22.7% in men and women respectively [3]. Furthermore, fewer than 50% of patients return to an independent lifestyle [4].

To diagnose a hip fracture in the Netherlands, radiological examinations (X-Rays) as described in the Dutch national guideline Proximal Femur fracture are employed [5]. In accordance with the guideline, fractures can be sub-classified according to the AO [6] and Garden standard [7]. A report is written by a radiologist on their findings with regards to the X-Rays. The classification of and the reporting on such examinations can be both time consuming and prone to inter-rater variability [8]. This, coupled with an expected increasing number of patients with proximal femur fractures [9], has raised the question of whether it would be valuable to automate these processes [10].

While previous work addressed automation of proximal femur fracture classification [10–12] and automated report generation [13–15] based on X-Rays, the clinical adaption of such systems is still hindered by several problems. One problem for *fracture classification* is the artificial

☆ This article belongs to Special issue: AIME 2019

* Corresponding author.
** Correspondence to: M. Nauta, University of Duisburg-Essen, Essen, Germany.
*E-mail addresses:* o.t.paalvast@student.utwente.nl (O. Paalvast), m.nauta@utwente.nl (M. Nauta), marion.koelle@uni-due.de (M. Koelle), j.geerdink@zgt.nl (J. Geerdink), o.vijlbrief@zgt.nl (O. Vijlbrief), h.hegeman@zgt.nl (J.H. Hegeman), christin.seifert@uni-due.de (C. Seifert).

https://doi.org/10.1016/j.artmed.2022.102281
Received 3 February 2021; Received in revised form 25 February 2022; Accepted 15 March 2022
Available online 26 March 2022

setting in which experiments are conducted. Both Gale et al. [10] and Jimenez-Sanchez et al. [11] rely on extensive pre-filtering of their datasets, which is time consuming and not feasible in practice.

We address this gap by presenting a model for fracture type sub-classification which requires no manual pre-processing.

Furthermore, *generated medical reports* as seen in related work are either reduced to simplistic templates as in work by Gale et al. [13] or are not validated by medical domain experts for their relevance and correctness in the clinical context, as in work by Gale et al. [10] and Jing et al. [15]. We address this issue by proposing a free-text based report generation model for X-Rays of proximal femur fractures. We evaluate the generated reports with both standard automated text evaluation metrics (BLEU and ROUGE scores [16,17]), and in qualitative and quantitative user studies to validate medical relevance. Specifically, our contributions are:

- We present an approach for automatic radiology report generation that requires minimal data pre-processing for both X-Ray images and radiology report texts yielding state-of-the-art performance for fracture classification and radiology report generation.
- We incorporate multi-view image information for classification. While the anteroposterior view is most commonly used in research, lateral views are of key clinical significance. Any view can be included in our architecture.
- We propose an automatic, semantic evaluation of the generated free text reports based on the medical conclusion in a report.
- In addition to the automatic evaluation, we perform qualitative and quantitative user studies with domain experts. While we found no significant difference in language quality between real and generated reports, our qualitative evaluation revealed concrete areas of improvement w.r.t. medical content in the generated reports.

Our overall approach to report generation relies on a language model trained on radiology reports, a fracture classification model trained on X-Ray images and a report generation model combining the output of both to generate textual radiology reports. Section 4 describes the architectures and interplay between those models. In addition to the technical evaluation (cf. Section 6), we present two user studies in Section 7, evaluating overall language quality (cf. Section 7.1) and medical content quality (cf. Section 7.2). We present a joint view on the implication of our results in Section 8 and conclude our work in Section 9.

## 2. Proximal femur fractures

Proximal femur fractures are predominantly seen in the elderly population [2], and are well-known to contribute to the morbidity and the mortality of this patient group [1]. This is further stressed by the observation that admitted patients suffered a cumulative mortality of 20% at four months after trauma [18,19]. The incidence of hip fractures is higher for women, who generally suffer more from osteoporotic bone degeneration and other factors linked to their higher average life expectancy [1,5]. Worldwide, the incidence of these types of fractures is estimated to increase from a yearly incidence in 1990 of 1.66 million to 6.26 million by 2050 [9]. In the Netherlands 13,000 elderly patients with hip fractures were recorded in 2012, which is expected to rise to 21,000 elderly patients by 2040 [20]. In 2010, hip fractures represented 53% of all healthcare expenses on osteoporotic-related fractures, where the mean total costs at 2-years follow up are over 19 thousand euros [21,22].

The Dutch national guideline “Proximale Femurfractuur” [5] describes the approaches to treat a patient with a suspected hip fracture. Initially, all patients are subjected to radiographic examination, with both anteroposterior (AP) and axial imaging directions, to classify the fracture according to the AO [6] and Garden standard [7]. As 3–4% of all patients harbour an occult fracture, these examinations may be inconclusive and thus needing further imaging in the form of CT or MR scan to be performed [5,23]. In addition to the identification of the fracture, the correct sub-classification of the fracture according to the AO standard has a significant impact on diagnosis, treatment and prognosis [8]. The classification is a process that is error-prone, as among residents and experts only up to 71% agreement was reached on the type of fracture [8]. This disagreement in combination with potential delays, incurred from having to resort to other imaging techniques, could lead to increased mortality [24,25], prolonged hospital stay [26] and compounding costs [27].

Based on the examinations the proximal femur fracture can be classified and grouped into two types of femoral neck fractures and three types of pertrochanteric fractures [5,6]: i) femoral neck fractures without displacement, ii) femoral neck fractures with displacement of the neck, iii) stable pertrochanteric fractures (AO type A1), iv) unstable pertrochanteric fractures (AO type A2), and v) unstable reversed type pertrochanteric fractures (AO type A3). Note that the subtrochanteric femur fracture is not considered here. The sub-classification of a fracture plays an important role in both choosing the correct operative treatment as well as having implications for the potential recovery of a patient [8]. The treatment options of these fractures generally fall into two categories; the so-called internal fixation (IF) option and the prosthetic option. In both cases surgical intervention is required, as a nonoperative treatment plan is rarely ever advised [5].

## 3. Related work

In this section we review work on classification of proximal fractures and automatic generation of radiology reports in natural language. Additionally, we provide an overview on user studies assessing the quality of generated reports.

### 3.1. Automatic proximal femur fracture classification

Transferring from standard computer vision tasks, CNN-based models have been applied to the task of classifying proximal femur fractures. Most notably, Gale et al. [10] reported an accuracy of 97% on the fracture versus no fracture task, comparable to expert performance, where over 53,000 images were used in total. In their work a three-part pipeline for identifying a Region of Interest (ROI), excluding non-AP images and classifying a fracture was employed. In comparable works an even greater emphasis was placed on this ROI identification and extraction, as the authors argued that a network should be optimised for a localisation loss as well as a classification loss [11,28]. These networks are trained using spatial transformers, where an affine transformation of an input image is used to automatically extract a specific part of the image [29].

In the works of Kazi et al. [28] and Jiménez-Sánchez et al. [11] a fracture was further subclassified into one of six types according to the AO classification standard for types A1–A3 and B1–B3. In both works the manual extraction of a ROI showed the most promising results, where an average six-class classification accuracy of 66% and 46% were reported respectively. In more recent work by Krogue et al. [30] an entirely different set of class labels were used and an accuracy of 90.4% was obtained. An effort to improve performance by Jiménez-Sánchez et al. [12] manifested in the grouping of all pertrochanteric (AO type A) and column (AO type B) fractures, which resulted in a better classification accuracy of 91%.

Limited test set sizes, differing class labels and different datasets make it impossible to directly compare these results. Additionally, manual preprocessing of the dataset is an integral component of the discussed related work. Here either manual selection of data or manual extraction of ROI's is necessary to attain published results. In this work hard selection or preprocessing of the data is not performed in an effort to reduce manual and subjective labelling of the dataset. The advantages and novelties of this approach are two-fold. First, by using data as-is a

realistic performance measure can be obtained with regards to images as they'd come available in clinical practice. Furthermore, the use of lateral views in addition to the standard AP views enables us to leverage information from both image sources in a committee voting approach to improve upon state-of-the-art proximal femur fracture classification.

### 3.2. Automated report generation

In addition to deriving whether an image contains a fracture or not, a report has to be written on these findings. These reports contain further information on the classified image, but can be cumbersome to write as many of these reports will be very repetitive and the writing of such reports is not the top priority of a radiologists' workflow [31]. Yet these reports are very important, as they serve as a communication tool of a summary of key information between doctors. In 2016, Xu et al. [32] created a model which automatically generates a caption based on an input image. The authors proposed an encoder-decoder architecture with a CNN as encoder, and an RNN decoder to generate text.

These ideas were extended to proximal femur fractures in work by Gale et al. [13]. They found that a physician was much more likely to accept a classification generated by a model if it was accompanied by a textual explanation in the form of a radiological report. Additionally, the ability to generate medically relevant textual descriptions of X-rays has the potential to standardise reporting on diagnostic imaging with only relevant information, to alleviate time-related burdens radiologists face, to reduce miss-classifications and consequently increase report quality and reduce costs. In their efforts to construct such a system, however, Gale et al. found it challenging to overcome the problems linked to the high variability in the content of radiological reports. To combat this a scaffold sentence was created in which several words were filled in by the model, which despite showing success was evidently a significant simplification of the original problem. A more realistic solution was offered by Wang et al. [14] who added multi-level attentions on a model initially trained solely for classification. They showed that it was possible to create accurate free-text radiological reports based on chest X-Rays. This is corroborated further by Han et al. [33], who showed similar success for reporting spine radiographs.

In this work the free-text based line of thinking is continued. The novelty introduced in this work is the direct encoding of input images based on free text radiology reports by using the proximal femur fracture classification model. This enables a tandem evaluation of the generated reports with the classification model. Additionally, this work is the first to use Dutch as a primary language for the report generation aspect in combination with the target field of proximal femur fracture classification.

### 3.3. Evaluating with users

There has been, to our knowledge, relatively little work on the topic of qualitative analysis of generated medical reports and image captioning in general. Jing et al. [15] mention qualitative analysis briefly, but also indicate that the authors themselves were the ones evaluating. Li et al. [34] approached non-domain experts as participants where each participant had to match a real report to the best option out of several generated reports. In their work to automatise medical text summarization MacAveny et al. [35] had their results analysed qualitatively by a single radiology expert. Recently, Li et al. [36] approached medical domain experts in their analysis but only asked them to rank several generated reports. These works all focused on chest X-ray based reports. On the topic of proximal femur fractures [13], experts were also recruited but were only subjected to a ranking question setup. Qualitative evaluations are far more common in other fields, were the open coding in grounded theory framework is one of the avenues used to evaluate the responses of domain experts [37,38]. We subject our method to user studies with medical domain experts assessing both, language quality and medical content of the generated reports. We apply inductive category development [39] to identify open issues and key motifs [40]. By presenting a thorough explanation of both our method as well as our evaluation we intend to inspire future work and serve as good practice for evaluations of Deep Learning based models in a clinical setting.

## 4. Approach

Given a case $C$ containing a set of radiology images taken for a certain patient, the goal is to (i) determine if the patient has suffered a fracture, and if so which sub-type it is, and (ii) automatically generate an accurate radiological report.

To generate a report, three machine learning models are combined as shown in Fig. 1. The classification model $f(\cdot)$ is used to obtain image embeddings. A language model trained on radiology reports outputs word embeddings as representation of radiology reports. The final report generation model is then trained to learn the connection between the input image embeddings and the word embeddings of the corresponding radiology report. During testing phase the report generation outputs a radiology report text solely based on the image (embeddings).

### 4.1. Fracture classification model

To avoid information leakage from the training data, we split our data set on a per case basis. This means, all images taken for one patient during a visit are either part of the training, the test or the validation set.

To classify fractures we train a model $f(\cdot)$ that takes $i$ as input and predicts a class label $\hat{y} = f(i)$, where $i \in C$ is a single X-ray image and the class label $\hat{y} \in$ {no fracture, column fracture, trochanter fracture}. Our classification task consists of two fracture classes and a no fracture class even though the clinical classification standard uses five fracture classification types. Similar to related work [12], we chose to aggregate fracture types with few training samples, and only distinguish column and trochanter fractures. The clinical impact of this decision is largely cost-related and not of influence on the outcome for the patient [5]. The corresponding ground truth class labels were obtained from the original radiological reports using the extraction described in Appendix A.3.

The decision for a case is obtained by aggregating the decisions for all images in $C$. More specifically, we calculate the mean posterior probabilities over all predictions for a single image and assign the label with the maximum average score. This approach also mimics the clinical decision making process closely; in this setting all images are used to reach a conclusion instead of finding a conclusion for each image separately.

The classification model learns a representation of the images. The latent representations in later layers also capture information about the type of fracture. More specifically, we use the features from the penultimate dense layer, which represents a 1024 dimensional encoding of an input image. These latent representations serve as input to the report generation model.

### 4.2. Language model

The language model is specifically trained on the language used in radiological reports and serves to guide the construction of generated reports. It is used to learn embeddings of words found in the reports such that they may be used as input to the report generation model. To train this language model, a separate language model dataset is used. This dataset consists of radiological reports of proximal femur fracture cases that are largely disjoint to the main dataset (cf. Section 4.3). This secondary dataset is used solely to train the language model.

### 4.3. Report generation model

The report generation model is trained with the word embeddings for

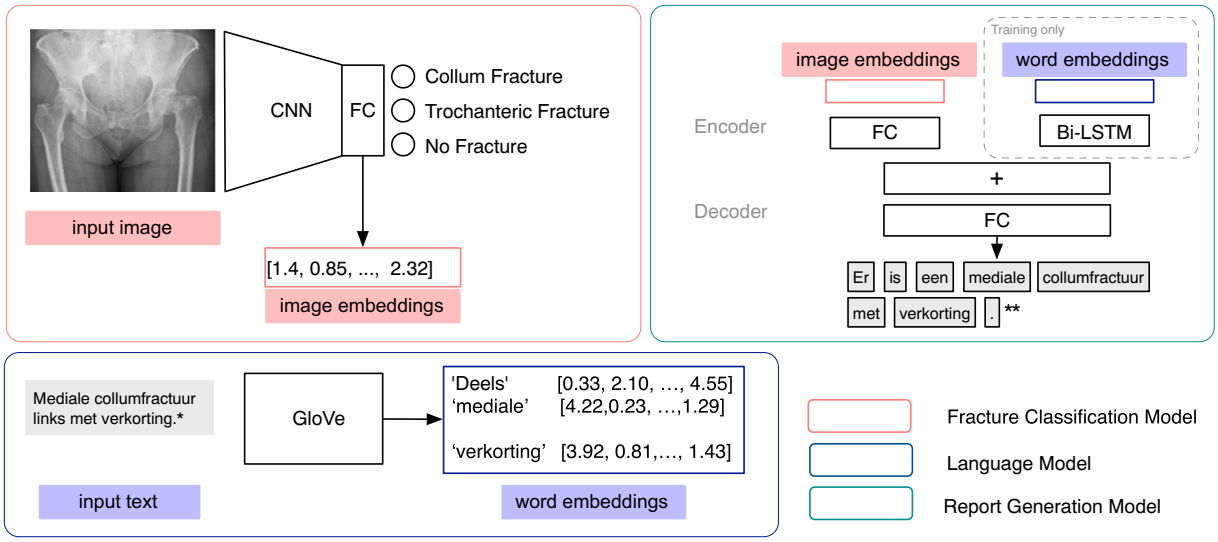



**Fig. 1.** Report generation model pipeline. The fracture classification model predicts the type of fracture for an image. The language model learns word embeddings for the associated radiology report. The report generation model is trained on image embeddings from the fracture classification model (FC - fully connected layer) and the word embeddings of the associated radiology report generated by the language model. During test time, the report generation model outputs a report solely based on the image embeddings. *Medial column fracture on the left side with shortening. **There is a medial column fracture with shortening.

a report and the image encoding for the corresponding image. The original, human-written report is used to compute a cross-entropy loss function. An overview of the model pipeline is shown in Fig. 1.

The report generation model uses an encoder-decoder architecture. The encoder obtains two inputs during training: the image encoding and the per word level embeddings. The image encodings are processed through a fully connected layer (FC), and combined with the Bi-LSTM output of the word embeddings using element-wise vector addition. The combined input is fed through a fully-connected layer consisting of 256 elements and into a softmax layer, which size corresponds to the number of words in the vocabulary. During testing, the word embeddings of the original report are not available. To bootstrap the language generation, the word embedding of the '<start>' token is input alongside the image embedding. The model outputs a single word at a time. A full sequence of words is created by maximising the joint probability of the predicted sentence:

$$logp\left(\mathbf{x}_{0:T}|I;\theta\right) = \sum_{t=0}^{T} p\left(\mathbf{x}_t|I,\mathbf{x}_{0:t-1};\theta\right) \tag{1}$$

where $I$ represents the image encoding and $\theta$ the model parameters. The variable $\mathbf{x}_{0:T}$ represents the words that have been generated at positions 0 through $T$. The first inference method is the greedy approach. In this approach the next word in a sequence is found as follows:

$$\mathbf{x}_t = \underset{w \in W}{max} p\left(\mathbf{x}_w|I,\mathbf{x}_{0:t-1};\theta\right) \tag{2}$$

where $w$ is a word in the vocabulary $W$. This method is called the greedy approach because it only optimizes for the next best step. For a time step $t$ only the single next word at $t+1$ is considered. A combination of words that individually might not have the highest probability is skipped even though their combination might have an ultimately greater joint probability. To account for this, the more computationally expensive beam search [41] has been proposed. At every time step $t$ a fixed number of candidates are evaluated. For each of these candidates an evaluation is made at time $t+1$ from which the sequence with the highest joint probability is selected.

## 5. Dataset

We apply our methods to data provided by the ZGT hospital in Hengelo/Almelo, The Netherlands. We obtain two datasets: a language model dataset with radiological reports on pelvic region fractures and our main data set with images and associated radiological reports. All reports were in Dutch.

### 5.1. Data extraction

Both datasets were extracted from the Picture Archiving & Communication System of the ZGT hospital. A complete description of the extraction criteria can be found in Appendix A.4. The reports for the language model dataset were sourced from a pool of radiological reports from the hospital data base. These reports were used because they were easily accessible and for the purposes of the language model dataset did not need to be checked for data quality. The corresponding SQL queries are reported in Appendix A.4. The reports of the language model dataset were subjected to the report extraction process as described in Appendix A.2. A large number of these reports used a different formatting and were discarded for this reason. In total over 28,000 reports were successfully parsed and added to the language model dataset.

### 5.2. Preprocessing

We downsize the images to a resolution of 300 × 300 pixels for faster training of the networks.[1] During training of the fracture classification network we transform the input images to artificially increase the amount of image data and their variance. All transformations are designed such that the expected region of interest around the femur head remains fully visible in the image.[2] We apply the following image transformations: (i) random rotation by up to 40°, (ii) random crops such that at least 80% of the original image remains, (iii) shearing by a random factor of up to 0.2, and (iv) horizontal flipping. Note that these image augmentation steps are only applied for training the fracture classification model, but not applied when training the report

[1] Preliminary experiments showed the same prediction accuracy compared to original image sizes.

[2] Confirmed by manually investigating random samples.

generation model.

All textual reports are preprocessed as follows. Any personal information is removed and multiple subsequent spaces are deleted. All non-alpha-numeric characters are removed. The sentence 'There is no fracture' (in Dutch) is used as the report for all no fracture cases. This is done to focus the text generation on the fracture characteristics. The reports used for the report generation are extended by adding a start token '<start>' and an end token '<end>'.

### 5.3. Dataset statistics

The dataset for training the language model contains 28,329 radiological reports. The main data set contains 4915 cases with their corresponding radiological reports and 11,606 images. Each case contains at least one image, with an average of $2.5 \pm 0.7$ images per case. Table 1 shows an overview of the class distribution. The class label, i.e., the fracture type, is available for each case. Thus, all images for one case have the same class label, independent of whether the fracture might be visible in the image or not. Most reports contain between 15 and 25 words; the distribution for the training data set is shown in Fig. 2. We split both datasets randomly into 70% training, 20% validation and 10% test data. For the main data set we use a case-based split (cf. Section 4.1).

## 6. Technical evaluation

We evaluate the fracture classification and report generation models separately using standard evaluation metrics. Results of the user studies are reported separately, in Section 7. We used Keras[3] to implement all models and the image transformations. We trained on a Linux system with Ubuntu 18.04 on an Intel Xeon E-2124 CPU, 32 GB RAM and a RTX 2080 GPU.

### 6.1. Model training

The *classification model* is based on the DenseNet architecture as used in related work [10]. Specifically, we use DenseNet-169 [42] pretrained on the ImageNet dataset [43]. We append two dense layers of 1024 units each with a ReLU activation function and a softmax layer. Both dense layers use a dropout rate of 0.5 [44]. We train all layers of our architecture using the Adam optimiser [45] with a learning rate of 0.00001 and the categorical cross-entropy loss function. Our batch size is 4 images. We train for 100 epochs, and apply early stopping of 7 epochs on the validation loss.

As *language model*, we trained a GloVe model [46] with a window size of 5 words, an embedding dimension of 100, a batch size of 4096 and a vocabulary size of 10,000. The model was trained for 100 epochs with a learning rate of 0.0001 and early stopping for the GloVe loss function.

To train the *report generation model* we truncated the reports to a maximum length of 40 words. This setting keeps the majority of the reports but does remove some of the very long outlier reports (cf. Fig. 2). Shorter reports are zero-padded and a per word level embedding for each report is obtained from the trained GloVe model. These 100 dimensional word embeddings are passed through a 128 dimensional Bi-LSTM layer resulting in a text embedding of size 256.

The image encoding of length 1024 is obtained from the fracture classification model by removing the final Dense layer, which results in a 1024 length encoding. This is then passed through a 256 fully connected layer in the encoder of the model.

**Table 1**
Overview of the datasets.

| | Type | # of cases | # of images | # of reports |
|---|---|---|---|---|
| Main dataset | No fracture | 2068 | 5068 | 2068 |
| | Column fracture | 1585 | 3606 | 1585 |
| | Trochanter fracture | 1262 | 2932 | 1262 |
| | Total | 4915 | 11,606 | 4915 |
| Language model dataset | Fracture reports | – | – | 28,329 |

[3] https://keras.io/.

### 6.2. Performance evaluation

For the fracture classification we measure accuracy, precision and recall on the test set. We use 5-fold cross-validation and report mean and standard deviation. In order to compare our results to related work, we also apply post-processing to evaluate the performance on the fracture vs. no-fracture task. In this post-processing step, the different fracture types are all considered to be 'fracture'.

To compare with related work, we report our results on a per-image basis, and additionally on a per case basis, since this is more clinically relevant (cf. Section 4.3). Finally, a so-called 'committee-voting' approach is applied where the information that multiple images belong to the same patient is leveraged. In this committee-voting evaluation approach the final classification is the average of all predicted model classifications, i.e. the raw prediction scores, for the images included in a case for a single patient.

The GloVe embedding model is trained until convergence. No specific metrics are reported.

We evaluate the report generation model using the standard text similarity scores BLEU-1, BLEU-2, BLEU-3, BLEU-4 [16], and ROUGE-L [17]. Additionally, we assess the semantic correctness w.r.t. to the final diagnoses by parsing the conclusions corresponding to no fracture, column fracture or trochanteric fracture from the generated reports and compare it to the ground truth. This parsing is described in detail in Appendix A.2. We report the percentage of agreement. Similarly to the classification model, we evaluate the report generation model on a per-patient basis using a committee voting approach by averaging the scores for all single predictions of a patient.

### 6.3. Results and discussion

Results for the evaluation of the *fracture classification model* show that the committee voting approach outperforms performance for single-image classification (cf. Table 2). This is due to the fact that the ground truth label is available on a per-case base, and not on a per-image level. A comparison of best case models shows that our approach is competitive with regards to related work (cf. Table 3). The classification model achieved an overall test set accuracy of $87 \pm 2\%$, or $90 \pm 2\%$ with committee voting. These results represent an improvement over the upper bound models reported in [28] and [11]. Additionally, our approach works on data that is less clean and relies on fewer assumptions regarding data distribution.

For the binary classification task our results are an improvement upon the results in [11], [28] and [12]. This improvement is realised without the extensive manual labelling necessary to create the ROI extraction pipelines these works propose. Compared to [10] the results are slightly worse, but it is unknown what effect their extensive dataset filtering performed has in their work.

Our results for the *report generation* are shown in Table 4. We used a simple default sentence for the no fracture cases, which the model could nearly always perfectly reconstruct. As this skews our results, we also report results only on the reports for the fracture cases.

Our BLEU scores are significantly lower than those reported in [13]. In their work, the authors use a template sentence and only asked the model to fill in the missing parts, resulting in a much easier task for their model. Additionally, it is not entirely clear which preprocessing steps

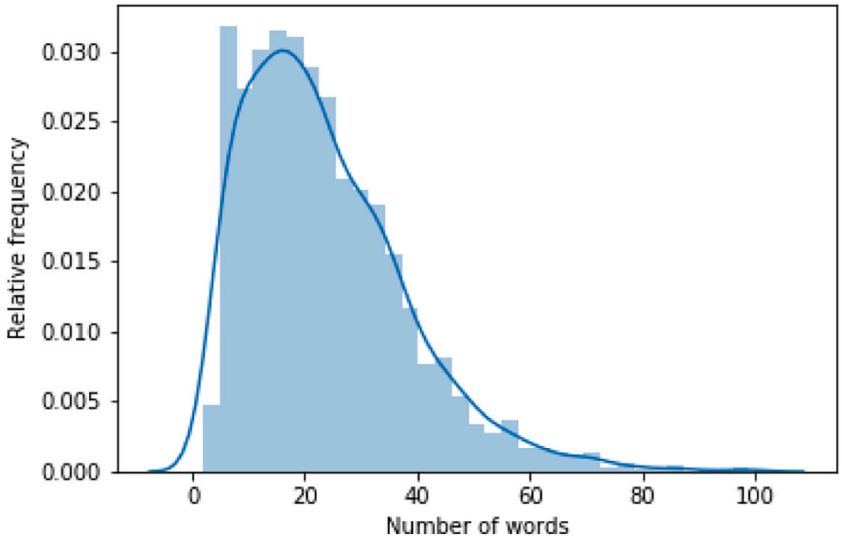



**Fig. 2.** Distribution of report lengths in the training data.

**Table 2**
Results for the four proposed experiments evaluating the fracture classification model. Reported are precision (prec.), recall (rec.), F1-score and accuracy (acc.) for the column-, trochanteric- and no fracture setup. Additionally, a fracture versus no fracture setup is reported upon.

| Class | Standard evaluation | | | | Committee voting | | | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. |
| Column | 0.86 ± 0.04 | 0.84 ± 0.03 | 0.85 ± 0.01 | 0.87 ± 0.02 | 0.90 ± 0.05 | 0.88 ± 0.02 | 0.89 ± 0.02 | 0.90 ± 0.02 |
| Troch. | 0.88 ± 0.05 | 0.77 ± 0.06 | 0.82 ± 0.03 | | 0.93 ± 0.04 | 0.82 ± 0.05 | 0.87 ± 0.03 | |
| No frac. | 0.89 ± 0.02 | 0.95 ± 0.02 | 0.92 ± 0.01 | | 0.90 ± 0.02 | 0.97 ± 0.02 | 0.93 ± 0.01 | |
| Frac. | 0.96 ± 0.01 | 0.90 ± 0.02 | 0.93 ± 0.01 | 0.93 ± 0.01 | 0.97 ± 0.01 | 0.92 ± 0.02 | 0.94 ± 0.01 | 0.94 ± 0.01 |
| No frac. | 0.89 ± 0.02 | 0.96 ± 0.02 | 0.92 ± 0.01 | | 0.90 ± 0.02 | 0.97 ± 0.02 | 0.93 ± 0.01 | |

were taken, i.e., whether the authors also do impose as little constraints on the original data as we do. Our results match the findings for the more comparable free-text based report generation [14]. Inspecting examples with high and low text similarity, we observe a repetition of key phrases such as 'Lateral column fracture on the left side' and 'Medial column fracture on the right side' (translated from Dutch) in the generated reports. Inspecting examples of generated reports (cf. Table 6), we observe that the BLEU and ROUGE metrics do not capture semantic similarity well, even when key findings are retained. This is in line with observations in related work [47,34]. For this reason, we additionally performed a semantic analysis of the reports and two user studies (cf. Section 7) to assess the impact on clinical decision making. For the semantic analysis, we compared the conclusion from the generated reports, i.e., the extracted fracture type, to the ground-truth, and we found a high agreement (cf. Table 5). Notably, these results are very similar to the performance of the fracture classification model (Table 2) indicating that the report generation model has learned the importance of correctly reporting the fracture type.

## 7. User studies

As a complementary evaluation, focusing more on the human perspective, we present two user studies. Namely, we evaluate the

**Table 4**
Text evaluation metrics for the report generation model on the entire data set (including no fracture cases), and on the fracture cases only, which require a more diverse description.

| Metric | Entire dataset | | Only fracture cases | |
|---|---|---|---|---|
| | Greedy | Beam search | Greedy | Beam search |
| BLEU-1 | 0.65 | 0.64 | 0.16 | 0.20 |
| BLEU-2 | 0.63 | 0.60 | 0.09 | 0.09 |
| BLEU-3 | 0.61 | 0.57 | 0.04 | 0.04 |
| BLEU-4 | 0.60 | 0.56 | 0.02 | 0.01 |
| ROUGE-L | 0.69 | 0.64 | 0.24 | 0.19 |

**Table 3**
Results compared to related work. The 2-class setup represents a fracture vs. no-fracture experimental setup. The 3-class setup is not identical across all works, but a general distinction between column fractures and trochanteric fractures is made. Additionally, it is reported whether automated or manual pre-processing in either annotation or selection of data is used. Reported are best case models.

| Setup | Author | F1-score | Accuracy | Pre-processing |
|---|---|---|---|---|
| 3-Class | Jimenez-Sanchez et al. [11] | 0.87 | – | Manual |
| | Kazi et al. [28] | 0.79 | 0.86 | Manual |
| | Jimenez-Sanchez et al. [12] | – | 0.91 | Manual |
| | **Our work** | 0.91 ± 0.04 | 0.90 ± 0.02 | Automated |
| 2-Class | Jimenez-Sanchez et al. [11] | 0.94 | – | Manual |
| | Kazi et al. [28] | 0.91 | 0.88 | Manual |
| | Jimenez-Sanchez et al. [12] | – | 0.94 | Manual |
| | Gale et al. [10] | 0.97 | 0.97 | Manual |
| | **Our work** | 0.94 ± 0.02 | 0.94 ± 0.01 | Automated |

**Table 5**
Evaluation of the report generation model where the agreement is measured between the fracture type extracted from the generated reports and the fracture type ground truth. The predicted labels are obtained by extracting the predicted fracture type (column, trochanteric or no fracture) from the generated reports. These are compared against the ground truth fracture type for the case corresponding to the report.

| Class type | Standard evaluation | | | | Committee voting | | | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | Acc. | F1 | Prec. | Rec. | Acc. | F1 |
| Column fracture | 0.91 | 0.87 | 0.89 | 0.89 | 0.92 | 0.90 | 0.92 | 0.91 |
| Trochanteric fracture | 0.87 | 0.80 | | 0.84 | 0.89 | 0.81 | | 0.85 |
| No fracture | 0.90 | 0.93 | | 0.91 | 0.91 | 0.95 | | 0.93 |

**Table 6**
Example reports with high and low ROUGE-L scores (R-L). Translated report from Dutch to English in italic.

| Original report | Generated report | R-L |
|---|---|---|
| Mediale collumfractuur rechts met een varusstand. Bekkenring is verder intact. Beide heupen vertonen beginnende coxarthrosis. Degeneratieve veranderingen lumbosacrale overang./*Medial column fracture on the right side with a varus position. Pelvic ring is intact. Both hips show initial coxarthrosis. Degenerative changes in the lumbosacral junction.* | Er is een collumfractuur rechts met trochanterhoogstand. Klein botfragment mediaal. Nog nagenoeg geen arthrose deformans. Bekkenring intact./*There is a column fracture on the right side with high trochanter presentation. Small bone fragment medially. No significant arthrosis deformans. Intact pelvic ring.* | 0.32 |
| collumfractuur links met inclavatiestand en slechts minimale dislocatie./*Column fracture on the left side with valgus impacted position and only minimal displacement.* | Kalkarm skelet. Collumfractuur links met verkorting. Uitgebreide vaatverkalkingen zichtbaar./*Skeleton low in calcium. Column fracture on the left side with shortening. Extensive calcifications of the vessels visible.* | 0.33 |
| Mediale collumfractuur links met verkorting/*Medial column fracture on the left side with shortening.* | Mediale collumfractuur links met verkorting/*Medial column fracture on the left side with shortening.* | 1.00 |
| Collumfractuur links met geringe verkorting/*Column fracture on the left side with slight shortening.* | Er is geen fractuur./*There is no fracture.* | 0.00 |
| Laterale collumfractuur rechts, met enige dislocatie./*Lateral column fracture on the right side, slight displacement.* | Normale botstructuur en kalkgehalte. Mediale collumfractuuur links. Verder fors degeneratieve veranderingen van de laaglumbale wervelkolom/*Normal bone structure and level of calcium. Medial column fracture on the left side. Significant degenerative alterations of the low lumbal spinal column.* | 0.10 |
| Er is geen fractuur/*There is no fracture* | Mediale collumfractuur links met verkorting/*Medial column fracture on the left side with shortening.* | 0.00 |

**Table 7**
Distribution of original and generated reports for trochanteric and column fractures in the full data set and in the sampled set used in Study I, respectively. The distributions show that the sample is representative for the full data set.

| | Full data set | | Samples set | |
|---|---|---|---|---|
| | Original reports | Generated reports | Original reports | Generated reports |
| Trochanteric fracture cases | 1585 (56%) | 168 (59%) | 56 (56%) | 114 (57%) |
| Column fracture cases | 1262 (44%) | 117 (41%) | 44 (56%) | 86 (43%) |

language quality of the generated reports with 90 participants having a background in Technical Medicine (Study I) and the medical content quality in a qualitative expert evaluation with 8 medical professionals (Study II). For these user studies, we selected the best performing model (cf. Section 6), i.e. the committee voting approach for the fracture classification (accuracy 90%, cf. Table 2) and the greedy inference strategy for the report generation (cf. Table 5). In the following, we detail on each of the two studies, before jointly discussing their key findings.

### 7.1. User study I: language quality

#### 7.1.1. Method

We employed an online questionnaire to quantitatively evaluate the general language quality of the generated reports: "The report's language quality is adequate for a medical report" (7-point Likert scale; from 1-strongly disagree to 7-strongly agree). In addition, we queried whether the report's level of detail was perceived sufficient to identify the medical condition, as well as its perceived provenance (human or machine generated). Order and selection of questions was made to mitigate bias, i.e., carry-over effects from perceived level of detail to language quality. The complete set of questions is comprised by Appendix B. Each participant was asked to answer these questions for six representatively sampled reports. In total, 300 reports were evaluated by on average 1.2 (SD = 0.4) participants each: each report received at least 1 rating. In the following we detail on how these reports were sampled.

#### 7.1.2. Sampled reports

In order to evaluate a wide variety of representative reports, while considering the participant's effort and motivation, we created a representative subset (sample) of reports. This allowed for each participant to evaluate exactly six reports that were selected in a pseudo-randomized fashion from in total 300 representative reports. This set consisted of 100 reports sampled from the original (i.e., human-generated) reports, and 200 machine-generated reports from a trained report generation model. To create a representative sample and ensure fairness, we employed further selection criteria. Namely, we ensured that the distribution of trochanteric and column fractures in both sampled subsets represents their distribution in the whole data set (see Table 7). In addition, we introduced upper and lower bounds for the sampled human-generated reports (which tend to be longer than the machine-generated ones) to prevent participants from inferring the report's provenance from their length.

#### 7.1.3. Participants

We recruited 63 Dutch-speaking students enrolled in a Technical Medicine master program via mailing lists and word of mouth (21 M, 42 F; age: 24 (SD = 2)). We made this recruitment choice, as their background in Technical Medicine made them likely to be comfortable with the terminology used in the reports: all of them had completed at least four weeks of clinical internships, and a subject-specific Bachelor's degree. This allowed them to make informed judgements on the reports' language quality. Nevertheless, we note that some reports might still use professional jargon not all students are familiar with. For this reason, we focus our analysis on their assessment of general language quality, and consider the question on the medical level of detail as additional background information.

### 7.2. User study II: medical content quality

#### 7.2.1. Method

In order to qualitatively evaluate the medical content quality of the generated reports, we conducted a complimentary user study with medical experts. In this study, human- and machine-generated reports were subjected to direct comparison: for each pair of reports (i.e., a human- and a machine-generated report describing the same case), the participants reported on how they perceived the reports' similarity in terms of level of detail, and impact on clinical decision making. Participants were asked to provide a concrete rating on a 7-point agreement scale (see Appendix B for details) and to justify their judgement in a free-text explanatory statement. Due to the limited number of participants in this study, and the resulting need for careful statistical interpretation, we focus our analysis on the qualitative statements instead. We analyze theme in-depth, using inductive category development [39], a form of open coding which allows one to identify open issues and key motifs [40].

#### 7.2.2. Participants

In order to gauge specialized quality criteria (here: medical content quality), expert knowledge of domain workers is essential. For this reason, we solicited feedback from 8 medical professionals, including three trauma surgeons, four resident trauma surgeons in training, and one nurse practitioner trauma surgery. All of them are medical healthcare providers at the ZGT Hospital. On average, our 8 participants (5 male, 3 female, 41 (SD = 10) years old) had 12 (SD = 7) years of (post-graduate) experience working in the medical field. Their unique background and experience of many years allows them to comment on practical relevance (e.g., of reported details), and the report's supposed impact on their treatment decision, which contributes to ecological validity. Yet, their expertise also makes them sparse and thus harder to recruit (e.g., compared to our participants in study I). We account for this lower number of participants by focusing on qualitative feedback.

#### 7.2.3. Sampled reports

We presented each participant with 4 cases, each including one human-generated and one machine-generated report. In total, we purposefully selected 32 pairs of reports to match the following criteria:

- one case includes a generated report of above average word count (>10 words).
- one case includes a generated report which is a misclassification.[4] The misclassified report was marked as such.
- the remaining two cases are randomly sampled.

This resulted in 4 cases, including exactly one misclassification, presented to each participant.

### 7.3. Results & discussion

This section jointly presents and discusses the results of Study I and Study II. We refer to the participants from Study I as *medical students*, and to the ones from Study II as *medical experts* with the individual experts denoted as E1, E2, …, E8 where quotes are given for illustration. For the qualitative analysis of Study II we furthermore report on occurrences of specific themes (denoted as n) in all expert assessments of the cases (denoted as capital N, 4 cases per participant, $N = 32$).

#### 7.3.1. Human or not? Converging language quality and level of detail

The results we obtained from Study I indicated an equivalency of the original (human-generated) and (machine-)generated reports in terms of their level of detail, language quality and humanness (see Fig. 3). After confirming that all collected data (questions Q1–Q3) is not normally distributed (Shapiro-Wilk test with $W$=0.90,$df$=1113,$p$<0.05), we used a Mann Whitney $U$ test to confirm that there is no significant difference in the medical students' assessment of the original and generated reports in terms of level of detail U($N\_o$=124,$N\_g$=247), $z$=14515.0, $p$>0.05, language quality U($N\_o$=124,$N\_g$=247), $z$=14299.0,$p$>0.05 and perceived humanness U($N\_o$=124,$N\_g$=247), $z$=14225.5,$p$>0.05. We note that this test statistic is chosen, because the original and generated reports presented to an individual medical student are sampled independently from each other, resulting in the testing conditions to be unpaired.

In Study II, original and generated reports were presented to the medical experts alongside each other. Experts were aware of the provenance (original/human or generated) and were specifically asked to comment on both reports' level of detail with regard to the treatment decision. Overall, they noted at least one type of missing information for 10 out of the 24 correctly classified reports (we detail on misclassifications in the subsequent section). In the majority of these cases the experts found details to be missing in both, the original and generated reports ($n = 5$). In three cases the original report lacked information present in the generated report ($n = 2$), and in two cases vice versa (n = 2): the generated report missed information contained in the original report.

These findings show that in the majority of the cases (14 out of 24), there was no important information missing in the correctly classified reports, meaning that both reports would lead to the same treatment decision, as illustrated by the subsequent case:

**Case RP-3**. **Original report**: We see a pertrochanteric fracture on the right. The pelvic ring itself shows no abnormalities. Mild coxarthrosis on the left.

**Generated report**: There is a pertrochanteric fracture on the right. Hip joints show coxarthrosis. Pelvic ring is intact.

**Expert assessment (E3)**: "[…] *Identical reports with same level of detail, although the treatment plan could be influenced by a more specific classification of the type of fracture; pertrochanteric A1/A2/A3, column displacement/no displacement.* […] *The reports are identical w.r.t. the fracture, so my treatment plan would be the same.*"

Additional details that the experts found relevant for their treatment decision and would have liked to see in the reports, included subclassification of fracture type ($n = 4$), displacement (n = 5), and arthrosis ($n = 3$), one mention was uncodeable (i.e., just referred to missing information).

We found it noteworthy that length and overall number of details comprised by a report did not necessarily imply all details relevant for the treatment decision were present, as illustrated by the subsequent example.

> RP-5/E5 1/ The original report is very limited. I miss the degree of displacement, but now w.r.t. shortening. 2/ I don't understand the remark in the generated report about the varus stand. The report is complete by also describing other structures.

This shows that machine learning and medical expertise need to go hand in hand. Only considering ROUGE and BLEU scores is insufficient for capturing the language quality and the level of detail required for the specific task. Our results show that domain knowledge is needed to assess the quality of a generated medical text.

#### 7.3.2. Data quality: ensuring sufficient quality and avoiding misclassification

As previously discussed, Study I showed that the reports generated by our model succeed in approximating the language quality and level of detail of the original reports. Yet, our results also show that in both, the original and in the generated reports there was a similarly high

[4] We consider a report a misclassification if the machine-generated report does not come to the same conclusion as the human-generated report.

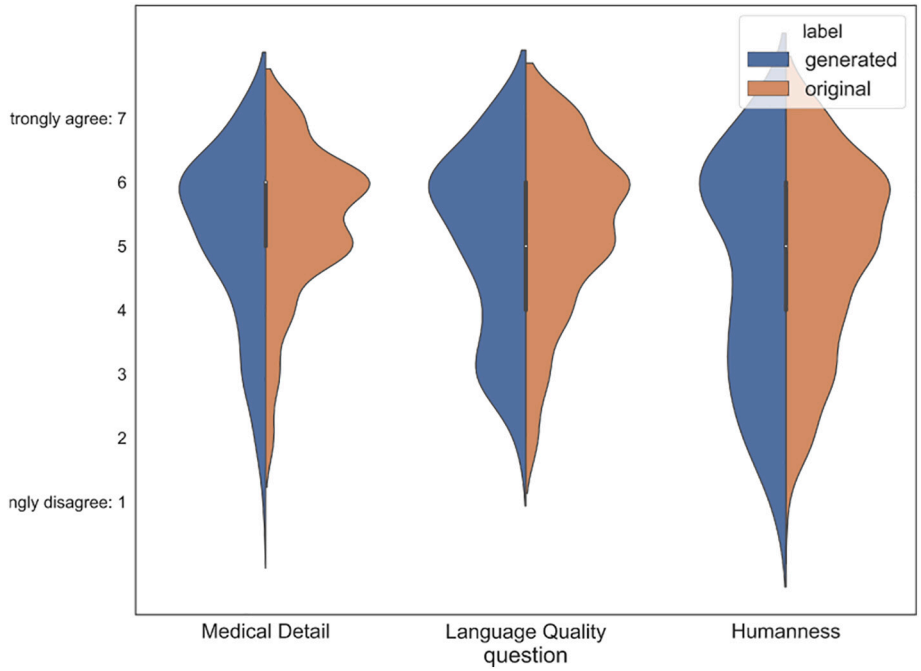



**Fig. 3.** Violin plots showing the distributions per question (Medical detail, Language quality or Humanness) per report source (original or generated).

variability of ratings (coefficient of variation of 0.26 and 0.31 for the original and generated reports respectively). This again illustrates how data quality in the original reports affects the quality of the generated reports – and potentially also how they are classified.

As a result of this consideration, Study II deliberately included one misclassified report per medical expert participant. A report was considered misclassified when the conclusion from the generated report did not match the ground truth of this case. Notably, all but one of the participants ($n = 7$) agreed that the misclassified generated report they were presented with was indeed misclassified. The misclassifications in the generated reports comprised incorrect lateralisation with respect to the fracture (five reports), and a misclassified type of fracture (four reports). In only three cases, inaccurate information in the generated reports was noted by the experts (n = 3). Interestingly, there was also a total of three *original* reports, where two of the medical experts (n = 3) identified misclassifications. E4 elaborates: "[…] *The diagnosis of the original report does not match my findings in the X-Ray.* […] *The diagnosis of the generated report is a better match, here a Garden classification would better describe the fracture*." In one other case, E3 found both the original and the generated report to be misclassifications. We followed up on this findings with an independent expert who verified that the original reports can be interpreted as misclassifications. This aligns with a phenomenon known from earlier research: for instance, Mast et al. describe a high inter-rater variability in classifying proximal femur fractures [48].

This finding has two implications: firstly, it might explain some erroneous classifications in the generated reports caused by incidental data issues. Secondly, and more importantly, it also highlights a challenge to be addressed by future work. The construction of a reliable ground truth data set (not covered by the present work) is as big a challenge as the construction of the model itself (covered by the present work). As directions for future work we suggest to pay particular attention to crafting a set of high-quality human-generated reports (similar to "textbook examples") that would then be used to further fine-tune the generated reports.

#### 7.3.3. How much detail?: impact of level of detail and options for extension

Analysing the quantitative results from Study I, we further found a strong positive correlation [49] between the ratings of medical detail and language quality (Pearson, $r(730)=.72, p<.001$) This, in addition to not finding a significant difference in language quality between the original and generated reports (cf. Section 7.2), suggests that the generated reports contain a level of medical detail that is considered sufficient by the raters. Language quality and humanness, and medical detail and humanness were only weakly to moderately correlated: Pearson, $r(730)=.49, p<.001$ and Pearson, $r(730)=.42, p<.001$. This suggests that medical detail and language quality are related to a greater extent than either of these is with the humanness of a text.

In Study II, participants were asked to comment on whether their treatment plan would be the same for both, the original and the generated report. Whether a report mentions or omits a specific detail (e.g., arthrosis) can potentially impact the treatment decision, as explained by E2: "*Depending on the patient characteristics the patient would be eligible for a KHP or THP based on either report. If the report would mention that there is coxarthrosis, my attention would be more so drawn to the possibility of the THP procedure. This is the case for the original report.*" Participants noted that details about the fracture ($n = 1$), presence and/or degree of displacement ($n = 4$), and presence of arthrosis (n = 1) are important for determining the treatment plan. These were found to be lacking in a total of five cases ($n = 5$), out of which half in the original reports ($n = 3$), and half in the generated reports (n = 3). In one case, both the original and the generated report lacked the required detail.

The results with regards to the level of detail carry three implications. The first one is related to the choice of using a free-text based approach for the generation of reports. In related work by Gale et al. [13] the level of detail was fixed by using a template sentence. In this template sentence the details that the participants in Study II indicated as necessary, e.g. save for the displacement, were not present. Our participants from study II mentioned what details they would have liked to see in the reports. Incorporating this information in a template sentence, or ensuring that those details are included in the training reports, opens up possibilities for further improvements.

Secondly, an explanation could be found in the model structure. The current method of inference, using a greedy approach, is rather naive and work by Bengio et al. [50] has already shown methods to improve upon this. Incorporation of a different inference scheme could also result in more detailed generated reports.

Thirdly, besides model improvements, the quality of the generated

reports depends on the dataset the model is trained on, as discussed in Section 7.3.2. The problems the participants encountered with the level of detail could stem from a lacking dataset. Anecdotal inspection of the dataset reveals that not all details deemed necessary are present in a majority of the training reports. This can partly be explained by the fact that a radiologist can judge an image on different characteristics than a trauma surgeon. Crafting a data set of high-quality reports with consistent language containing all important details could be used to further fine-tune the report generation model.

#### 7.3.4. What next?: concluding reports with a summary

Four of the medical experts (E3, E4, E5, E7) in Study II suggested that the addition of a concluding sub-classification based on the Garden or AO classification standard to the generated report would be valuable [5,6]. E4 suggests "*a Garden classification could describe the fracture even better*" (translated from Dutch).

Such conclusions could either be sourced from our classification model, the parsed conclusion from the report generation model, or be created using a summarization model. In previous work, this has already been successfully undertaken, e.g., by Marchawala et al. [51] on tasks as challenging as medical report summarization, which suggests that this is a viable path for future work.

## 8. Implications and outlook

We presented a competitive approach for the automated (sub-)classification of proximal femur fractures and free-text report generation model based on X-ray images. Even though the current efforts are successful, several improvements are possible. Both the classification model and the report generation model are trained and evaluated on single images. This does not realistically represent the clinical setting in which multiple images are taken for a single patient. It also introduces a problem where the original report describes findings over all images in a single case which might not be visible in each individual image. When training the report generation model the description is now taken as if it applies to every single image, which could lead to nonsensical results where details on both femurs are described in a lateral aspect image. To resolve this, a multi-input model could be considered where the input consists of all images taken for a single case to avoid mismatching descriptions and images. To some extent, this was implemented with the committee voting rule but a dedicated multi-input model would fully account for the specific multi-image nature of X-ray cases.

Prior to any clinical implementation we recommend the validation of the found results using an externally sourced dataset. Ideally this dataset is sourced from the most recent cases reported in an unrelated hospital so that this evaluation has the potential to show the actual clinical performance.

A promising addition to the current model architecture could be to incorporate an attention mechanism. In comparable report generation tasks, Xu et al. [32] and Wang et al. [14] incorporated an attention mechanism to improve their training performance.

## 9. Conclusion

In this work we presented a state-of-the-art approach for the automated (sub-)classification of proximal femur fractures from multi-view X-ray images. Additionally, we presented a free-text based report generation model based on X-ray images, obtaining competitive results in BLEU and ROUGE scores. The report generation model seems to lack in the reported text evaluation measures because of the free-text approach, but a confident agreement between the report generation conclusions and medical ground truths is shown. In contrast to existing work, both models require minimal data pre-processing for the included images and reports respectively. Thus, the assumption that data has to be cleaned and filtered to achieve state-of-the-art performance is substantially relaxed. We also successfully trained a language model underlying the report generation model for a language (Dutch) in which no relevant pretrained language model is available, thereby showing the applicability of our work in a domain and a language in which pretrained language models are not available.

Furthermore, we performed quantitative and qualitative user studies with medical domain experts. Our quantitative evaluation found no significant difference in language quality between the generated reports and the ground truth reports. Additionally, a high correlation was found between medical detail and language quality. The qualitative evaluation revealed concrete areas of improvements regarding the medical content of the generated reports. The medical domain experts involved remarked medical equivalency in several cases and noted occasional improvements of a generated report over the original. With such a model the development of a clinical application becomes feasible.

## Research data statement

Inquiries regarding access to the dataset used to train the classification model can be sent to: wetenschapsbureau@zgt.nl. To encourage participation and safeguard the sensitive nature of the questions asked in the two user studies the raw data for their responses is confidential. The code used for this paper will be made publicly available as supplementary material upon acceptance.

## Declaration of competing interest

One of the authors participated in the Medical Content Quality user evaluation (Study II). This author was not involved in the analysis reported in Section 7. There are no further conflicts of interest to report and no funding was received for this research.

## Appendix A. Dataset description

The dataset used for the caption generation model consists of 2000 fracture reports. Of these 81.8% had a collum fracture of some kind, with the rest being trochanteric fractures. The reports were provided in a.CSV format which included a number of fields:

- A unique identification number.
- The original report including HTML make up.
- A description tag added for the type of examination.
- An indication field
- A conclusion field
- A code field for registering the type of fracture
- A room description field corresponding to an X-ray examination room in the hospital.

The description tag for the type of examination, the code field for the type of fracture and the room description field are not used. For all reports the identification number, indication and the report field were filled in. The conclusion field serves as a means to structure reports by offering a suggestive

box in the electronic patient register system. This field is not filled in for 18.15% of the included reports.

An additional 2072 no fracture cases were included in the dataset as well. For these cases no reports were available so a standard sentence was included for all of the no fracture cases. This sentence read: 'Er is geen fractuur.'.

### A.1. The 'report' field

In order to train a caption generation model the raw reports found in the field containing the original report have to be preprocessed into a usable format. The radiological reports do seem to follow a general structure. A report contains an average of $32 \pm 19$ words and can be characterised by the following components:

1. A header indicating the start of the medical history section.
2. A section containing information on the medical history of a patient. This section usually contains medically relevant prior events or the medication that is being taken. It is used to indicate what the question is at hand and what transpired prior to the patients' arrival at the hospital.
3. A header indicating the start of the radiological report.
4. A section containing the radiological report on a taken series of X-ray images. This section describes the radiologists' findings and can include details on the entire pelvic area, the quality of the images and referrals to earlier examinations. If a fracture is present this is indicated alongside a localisation for this fracture.
5. A header for the conclusion of the radiological report.
6. A section containing the conclusion the radiologist reached in their report. This is generally a short sentence describing the type of fracture and its localisation.

The headers for the different components of the report don't follow a standard format. In the some 2000 reports used for the caption generation model over 50 different variations were found for the radiological report header alone. A number of different variations were found for the medical history header as well. The conclusion header is not present in an unknown number of reports. As mentioned before, the conclusion field, which is different from the conclusion header in the report field, is not always filled in. It is not necessarily true that in those cases a conclusion header and section are included in the report.

In addition to the disparities found for the report field, the report texts themselves show great inter- and intraoperator variance in terms of the language used. Out of a unique vocabulary of 2205 different words only 659 occurred five or more times. To alleviate the problems that such a sparsity of word usage could bring a word level embedding model is proposed to serve as a proxy for the input for the caption generation model.

### A.2. Parsing the report field

Given the found disparities in the section headers and the general structure of the radiological reports, it is expected that an automatic way to parse the reports will be hard to develop. This notion is further supported by work of Pathak et al. [52] who found the automatic parsing of such reports to be a challenging task. They reported results based on a per-word labelled dataset, which is simply unfeasible to obtain for this setting. As a result, a semi-automatic filtering method is proposed to extract the relevant information from the reports instead.

The filtering is done by using a manually constructed dictionary in addition to a set of regex-based filters. In an iterative approach the headers for the radiological report are parsed from the reports by running the corpus against the existing dictionary for matches. Any reports that fall through are then examined and the relevant header sections are added to the dictionary to make it possible to parse these reports. If a report is matched through a header from the dictionary, a regex based filter is applied to then extract the report text in the section that follows.

### A.3. Parsing the report conclusion

Having parsed a report a conclusion is then parsed from the result. This conclusion is one of two options; a collum fracture case or a trochanteric fracture case. Through a set of regex filters the fracture type is extracted from the original report. A manual dictionary is then used to map from the different fracture types to one of the conclusions.

### A.4. GloVe word embedding

There are generally two ways in which a word embedding can be constructed. The first method is called the one-hot encoding and this represents a simple mapping from an index to a word. In this scenario the word 'fractuur' could be assigned a number 18 for example. This method is very simple and quick to apply, but suffers from severe sparsity as the vocabulary size increases. Additionally, it doesn't account for the context of a certain word.

The alternative to the one-hot encoding method is the word embedding approach. A word embedding represents a vector space encoding of a single word. Instead of being represented by a single number, the word 'fractuur' is now transformed into a vector of numbers of a specified encoding dimension size. This has the primary advantage that an encoding can now capture information about a word's context as well. To illustrate this, it would now be possible to determine how 'close' two words are. In the context of radiological reports the words 'fractuur' and 'collum' are probably a lot more likely to appear in each others' context than the words 'fractuur' and 'wervelkolom'. This context knowledge is crucial to the caption generation step.

Several different algorithms exist for the construction of a model that can create these word encodings. One of these algorithms is the GloVe model. GloVe, short for Global Vectors, is a text embedding algorithm that takes into account both the context of words as well as their global statistics in the given corpus [46]. Having trained such a model, a word from a report can be transformed into an embedding which can then be fed into the caption generation model. The following SQL queries were applied to obtain this data:

- SELECT * FROM ZGT_RONTGEN WHERE (verslag LIKE '%fractuur%' OR verslag LIKE '%fraktuur%') AND verslag LIKE '%collum%' AND (omschr LIKE '%Bekken+heup beiderzijds%' OR omschr LIKE '%Bekken+heup links%' OR omschr LIKE '%Bekken+heup rechts%')

This resulted in 7595 reports.

- SELECT * FROM ZGT_RONTGEN WHERE (verslag LIKE '%fractuur%' OR verslag LIKE '%fraktuur%') AND verslag LIKE '%collum%' AND (omschr LIKE '%Bekken%')

  This resulted in 13,583 reports.

- SELECT * FROM ZGT_RONTGEN WHERE (verslag LIKE '%fractuur%' OR verslag LIKE '%fraktuur%') AND verslag LIKE '%collum%'

  This resulted in 23,389 reports.

## Appendix B. Evaluation surveys

*Language Quality*

*This part is repeated for all presented texts.*

1. **The report's level of detail is sufficient to identify the medical condition.**
   **strongly disagree** □—□—□—□—□—□—□ **strongly agree**
2. **The report's language quality is adequate for a medical report.**
   **strongly disagree** □—□—□—□—□—□—□ **strongly agree**
3. **The report is written by a human.**
   **strongly disagree** □—□—□—□—□—□—□ **strongly agree**

*Feedback and Comments*

4. **If you have feedback or comments for this study, please indicate it below.**

   _____

*Report Quality*

*This part is repeated for all presented texts.*

1. **The details I usually consider for my treatment decision are the same in both reports.**
   **strongly disagree** □—□—□—□—□—□—□ **strongly agree**
2. **My treatment decision would be the same for both reports.**
   **strongly disagree** □—□—□—□—□—□—□ **strongly agree**
3. **Please explain both of the above scores you provided.**

   _____

*Feedback and Comments*

4. **If you have feedback or comments for this study, please indicate it below.**

   _____

## References


[1] Dhanwal DK, Dennison EM, Harvey NC, Cooper C. Epidemiology of hip fracture: worldwide geographic variation. Indian J Orthop 2011;45(1):15–22. https://doi.org/10.4103/0019-5413.73656. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004072/.

[2] Brauer CA, Coca-Perraillon M, Cutler DM, Rosen AB. Incidence and mortality of hip fractures in the United States. JAMA 2009;302(14):1573–9. https://doi.org/10.1001/jama.2009.1462. https://jamanetwork-com.ezproxy2.utwente.nl/journals/jama/fullarticle/184708.

[3] Oden A, Dawson A, Dere W, Johnell O, Jonsson B, Kanis JA. Lifetime risk of hip fractures is underestimated. Osteoporos Int 1998;8(6):599–603. https://doi.org/10.1007/s001980050105.

[4] Morrison RS. The medical Consultant's role in caring for patients with hip fracture. Ann Intern Med 1998;128(12_Part_1):1010. https://doi.org/10.7326/0003-4819-128-12_Part_1-199806150-00010. http://annals.org/article.aspx?doi=10.7326/0003-4819-128-12_Part_1-199806150-00010.

[5] Specialisten FM. Aanvullend onderzoek proximale femurfractuur - Richtlijn - Richtlijnendatabase. https://richtlijnendatabase.nl/richtlijn/proximale_femurfracturen/diagnostiek_en_classificatie_proximale_femurfractuur/aanvullend_onderzoek_proximale_femurfractuur.html; 2016.

[6] AO. AO/OTA fracture and dislocation classification compendium-2018. https://classification.aoeducation.org/; 2018.

[7] Jin W-J, Dai L-Y, Cui Y-M, Zhou Q, Jiang L-S, Lu H. Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. Injury 2005;36(7):858–61. https://doi.org/10.1016/j.injury.2005.02.005. http://www.sciencedirect.com/science/article/pii/S0020138305000471.

[8] van Embden D, Rhemrev SJ, Meylaerts SAG, Roukema GR. The comparison of two classifications for trochanteric femur fractures: the AO/ASIF classification and the Jensen classification. Injury 2010;41(4):377–81. https://doi.org/10.1016/j.injury.2009.10.007. http://www.sciencedirect.com/science/article/pii/S0020138309005294.

[9] Cooper C, Campion G, Melton LJ. Hip fractures in the elderly: a world-wide projection. Osteoporos Int 1992;2(6):285–9.

[10] Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. 1711.06504 [cs, stat]ArXiv: 1711.06504. arXiv; 2017. http://arxiv.org/abs/1711.06504.

[11] Jiménez-Sánchez A, Kazi A, Albarqouni S, Kirchhoff S, Sträter A, Biberthaler P, Mateus D, Navab N. Weakly-supervised localization and classification of proximal femur fractures. 1809.10692 [cs]ArXiv: 1809.10692. arXiv; 2018. http://arxiv.org/abs/1809.10692.

[12] Jiménez-Sánchez A, Kazi A, Albarqouni S, Kirchhoff C, Biberthaler P, Navab N, Kirchhoff S, Mateus D. Precise proximal femur fracture classification for interactive training and surgical planning. Int J Comput Assist Radiol Surg 2020;15(5):847–57.

[13] Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing radiologist-quality reports for interpretable deep learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019. p. 1275–9. https://doi.org/10.1109/ISBI.2019.8759236.

[14] Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 9049–58.

[15] Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th annual meeting of the association for computational linguistics. 1; 2018. p. 2577–86. Long Papers.

[16] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics - ACL '02, Association for Computational Linguistics, Philadelphia, Pennsylvania; 2001. p. 311. https://doi.org/10.3115/1073083.1073135. http://portal.acm.org/citation.cfm?doid=1073083.1073135.

[17] Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74–81.

[18] van Balen R, Steyerberg EW, Polder JJ, Ribbers TL, Habbema JD, Cools HJ. Hip fracture in elderly patients: outcomes for function, quality of life, and type of residence. Clin Orthop Relat Res 2001;390:232–43.

[19] Folbert E, Hegeman J, Vermeer M, Regtuijt E, Velde D, Duis H, Slaets J. Improved 1-year mortality in elderly patients with a hip fracture following integrated orthogeriatric treatment. Osteoporos Int 2016;28. https://doi.org/10.1007/s00198-016-3711-7.

[20] Zorg LNA. Factsheet Heupfracturen. 2012.

[21] Zielinski SM, Bouwmans CA, Heetveld MJ, Bhandari M, Patka P, Biert J, Edwards MJ, Frolke JP, Al E, Lieshout EMv, Kampen Av. The societal costs of femoral neck fracture patients treated with internal fixation. Accepted: 2015-02-11T22:34:02Z885; 2014. https://repository.ubn.ru.nl/handle/2066/137078.

[22] Lötters FJB, van den Bergh JP, de Vries F, Mölken MPMHRutten-van. Current and future incidence and costs of osteoporosis-related fractures in the Netherlands: combining claims data with BMD measurements 2016;98(3):235–43. https://doi.org/10.1007/s00223-015-0089-z.

[23] Cannon J, Silvestri S, Munro M. Imaging choices in occult hip fracture. J Emerg Med 2009;37(2):144–52. https://doi.org/10.1016/j.jemermed.2007.12.039. http://www.sciencedirect.com/science/article/pii/S0736467908002199.

[24] Shiga T, Wajima Z, Ohe Y. Is operative delay associated with increased mortality of hip fracture patients? Systematic review, meta-analysis, and meta-regression. Can J Anesth 2008;55(3):146. https://doi.org/10.1007/BF03016088.

[25] Zuckerman JD, Skovron ML, Koval KJ, Aharonoff G, Frankel VH. Postoperative complications and mortality associated with operative delay in older patients who have a fracture of the hip. JBJS 1995;77(10):1551. https://journals.lww.com/jbjsjournal/Abstract/1995/10000/Postoperative_complications_and_mortality.10.aspx.

[26] Simunovic N, Devereaux P, Bhandari M. Surgery for hip fractures: does surgical delay affect outcomes? Indian J Orthop 2011;45(1):27–32. https://doi.org/10.4103/0019-5413.73660. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004074/.

[27] Shabat S, Heller E, Mann G, Gepstein R, Fredman B, Nyska M. Economic consequences of operative delay for hip fractures in a non-profit institution. Orthopedics 2003;26(12):1197–9. discussion 1199.

[28] Kazi A, Albarqouni S, Sanchez AJ, Kirchhoff S, Biberthaler P, Navab N, Mateus D. Automatic classification of proximal femur fractures based on attention models. In: Wang Q, Shi Y, Suk H-I, Suzuki K, editors. Machine learning in medical imaging, lecture notes in computer science. Springer International Publishing; 2017. p. 70–8.

[29] Spatial transformer networks. In: Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K, Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in neural information processing systems. 28. Curran Associates, Inc.; 2015. p. 2017–25. https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

[30] Krogue J. Automatic hip fracture identification and functional subclassification with deep learning, library catalog. www.groundai.com, https://www.groundai.com/project/automatic-hip-fracture-identification-and-functional-subclassification-with-deep-learning/1; 2020.

[31] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–10. https://doi.org/10.1038/s41568-018-0016-5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/.

[32] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: Bach F, Blei D, editors. Proceedings of the 32nd international conference on machine learning, Vol. 37 of proceedings of machine learning research. Lille, France: PMLR; 2015. p. 2048–57. http://proceedings.mlr.press/v37/xuc15.html.

[33] Han Z, Wei B, Leung S, Chung J, Li S. Towards automatic report generation in spine radiology using weakly supervised framework. series Title: Lecture Notes in Computer Science. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical image computing and computer assisted intervention – MICCAI 2018. 11073. Cham: Springer International Publishing; 2018. p. 185–93. https://doi.org/10.1007/978-3-030-00937-3_22. http://link.springer.com/10.1007/978-3-030-00937-3_22.

[34] Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in neural information processing systems. 31. Curran Associates, Inc.; 2018. p. 1530–40. http://papers.nips.cc/paper/7426-hybrid-retrieval-generation-reinforced-agent-for-medical-image-report-generation.pdf.

[35] MacAvaney S, Sotudeh S, Cohan A, Goharian N, Talati I, Filice RW. Ontology-aware clinical abstractive summarization. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, SIGIR'19. New York, NY, USA: Association for Computing Machinery; 2019. p. 1013–6. https://doi.org/10.1145/3331184.3331319.

[36] Li M, Wang F, Chang X, Liang X. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. arXiv; 2020. http://arxiv.org/abs/2006.03744.

[37] Vollstedt M. An introduction to grounded theory with a special focus on axial coding and the coding paradigm. 2019. https://doi.org/10.1007/978-3-030-15636-7_4. https://www.researchgate.net/publication/332700157_An_Introduction_to_Grounded_Theory_with_a_Special_Focus_on_Axial_Coding_and_the_Coding_Paradigm.

[38] Tie YChun, Birks M, Francis K. Grounded theory research: a design framework for novice researchers7. SAGE Open Medicine; 2019. https://doi.org/10.1177/2050312118822927. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318722/.

[39] Mayring P. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt; 2014.

[40] Boehm A. Chapter 5.13 "theoretical coding: text analysis in grounded theory"1. Sage; 2004. Long Papers.

[41] Freitag M, Al-Onaizan Y. Beam search strategies for neural machine translation. 1702.01806. In: Proceedings of the first workshop on neural machine translation. ArXiv; 2017. p. 56–60. https://doi.org/10.18653/v1/W17-3207. ttp://arxiv.org/abs/1702.01806.

[42] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv; 2016. http://arxiv.org/abs/1608.06993.

[43] Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I, editors. Artificial neural networks and machine learning – ICANN 2018. Cham: Springer International Publishing; 2018. p. 270–9.

[44] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(56):1929–58. http://jmlr.org/papers/v15/srivastava14a.html.

[45] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings; 2015. http://arxiv.org/abs/1412.6980.

[46] Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532–43. https://doi.org/10.3115/v1/D14-1162. http://aclweb.org/anthology/D14-1162.

[47] Lu S, Zhu Y, Zhang W, Wang J, Yu Y. Neural text generation: past, present and beyond. arXiv; 2018. http://arxiv.org/abs/1803.07133.

[48] Mast NH, Impellizzeri F, Keller S, Leunig M. Reliability and agreement of measures used in radiographic evaluation of the adult hip. Clin Orthop Relat Res 2011;469(1):188–99. https://doi.org/10.1007/s11999-010-1447-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3008883/.

[49] Evans JD. In: Straightforward statistics for the behavioral sciences. Belmont, CA, US: Thomson Brooks/Cole Publishing Co; 1996. p. 600. xxii.

[50] Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduled sampling for sequence prediction with recurrent neural networks. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in neural information processing

systems. 28. Curran Associates, Inc.; 2015. p. 1171–9. https://proceedings.neurips.cc/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

[51] Marchawala A, Patel P, Thaker KParesh, Gunjal H, Nagrecha A, Mohammed S. Text summarization and classification of clinical discharge summaries using deep learning. TechRxiv; 2020. https://doi.org/10.36227/techrxiv.12059019.v1. https://www.techrxiv.org/articles/preprint/Text_Summarization_and_Classification_of_Clinical_Discharge_Summaries_using_Deep_Learning/12059019.

[52] Pathak S, van Rossen J, Vijlbrief O, Geerdink J, Seifert C, van Keulen M. Post-structuring radiology reports of breast cancer patients for clinical quality assurance. IEEE/ACM Trans Comput Biol Bioinform May 2019. https://doi.org/10.1109/TCBB.2019.2914678.