Examples of impression sections of the radiology reports for different annotation and classification results. Shown are the original Dutch text, the English translation, the category of actionable finding (1, 2, 3, or No) annotated by each of the two annotators, the final reference for actionable finding (Yes or No), and the probability of an actionable finding according to the classifier.

Example 1 – Annotator disagreement (Category 3 vs. No category)

Geen argumenten voor lymfadenopathieën metastasen wel enkele aspecifieke longhaarden.
No arguments for lymphadenopathies metastases, but some non-specific lung foci.
Annotator 1: No; Annotator 2: Category 3 Reference annotation: No Classifier probability: 0.044
Remarks: Annotator 1 interpreted the non-specific lung foci as not actionable. Annotator 2 interpreted the term non-specific as if it could be metastasis and hence qualified it as a category 3 finding. During the consensus meeting it was agreed that the formulation was too vague to qualify it as a category 3 finding.

Example 2: Annotator disagreement (Category 2 vs. No category)

Residuele pneumothorax links. Slechte beluchting van de rechterlong, DD atelectase,infiltraat of een combinatie van beiden.
Residual pneumothorax left. Poor aeration of the right lung, DD atelectasis, infiltrate or a combination of both.
Annotator 1: Category 2; Annotator 2: No Reference annotation: No Classifier probability: 0.853
Remarks: Annotator 1 interpreted the pneumothorax as a category 2 finding. Annotator 2 interpreted the term “residual” as if the pneumothorax was already present, and thus as a non-actionable finding. This latter interpretation became common sense during the consensus meeting.

Example 3: Reference No actionable finding, classifier low probability

Status na artrodese bovenste spronggewricht, geconsolideerd.
Status after arthrodesis upper tarsus, consolidated.
Annotator 1: No; Annotator 2: No Reference annotation: No Classifier probability: 0.046
Remarks: Both annotators agreed to classify this finding as not actionable; the classifier generated a concordant result, i.e. a low probability.

Example 4: Reference Actionable finding, classifier moderate probability

Infiltratieve afwijking in de linkeronderkwab.
Infiltrative abnormality in the left lower lobe.
Annotator 1: Category 2; Annotator 2: Category 2 Reference annotation: Yes Classifier probability: 0.312
Remarks: Both annotators agreed to classify this finding as category 2; the classifier generated a equivocal result, i.e. an intermediate probability.

Example 5: Reference Actionable finding, classifier high probability

Forse hoeveelheid pleuraal vocht rechts, toegenomen. Ontstaan van verdichting retrocardiaal.
Significant amount of pleural fluid on the right, increased. Onset of retrocardiac compaction.
Annotator 1: Category 2; Annotator 2: Category 2 Reference annotation: Yes Classifier probability: 0.824
Remarks: Both annotators agreed to classify this finding as category 2; the classifier generated a concordant result, i.e. a high probability.

Scatter plots of word frequencies in radiology reports

The actionable finding classifier determines the probability that an actionable finding is present. To obtain a dichotomous classification (presence or absence of an actionable finding), we set the probability threshold to the probability that maximizes the F-score (harmonic mean of precision and recall: 2 * precision * recall / (precision + recall)) in the test set (maximum F-score 0.725, threshold 0.466). In the radiology reports with actionable findings according to the reference, we compared the relative frequencies of words in the impression section of the reports that were incorrectly classified (false negatives) with those that were correctly classified (true positives). Similarly, in the reports without actionable findings, we compared the relative frequencies in the impression section of the incorrectly classified reports (false-positives) and the correctly classified reports (true negatives). Figure S1 shows the relative frequencies for words in the false-negative and true-positive reports. Words on the diagonal have the same relative frequency in both report sets, while words above the diagonal are overrepresented in the false-negative reports and words below the diagonal are underrepresented (or overrepresented in the true-positive reports). The words on the diagonal with high relative frequencies are stopwords (“van”, “de”, “in” (English “of”, “the”, “in”)). An overrepresented word in the false-negative reports is “normale” (“normal”), possibly lowering the classifier’s probability of an actionable finding.

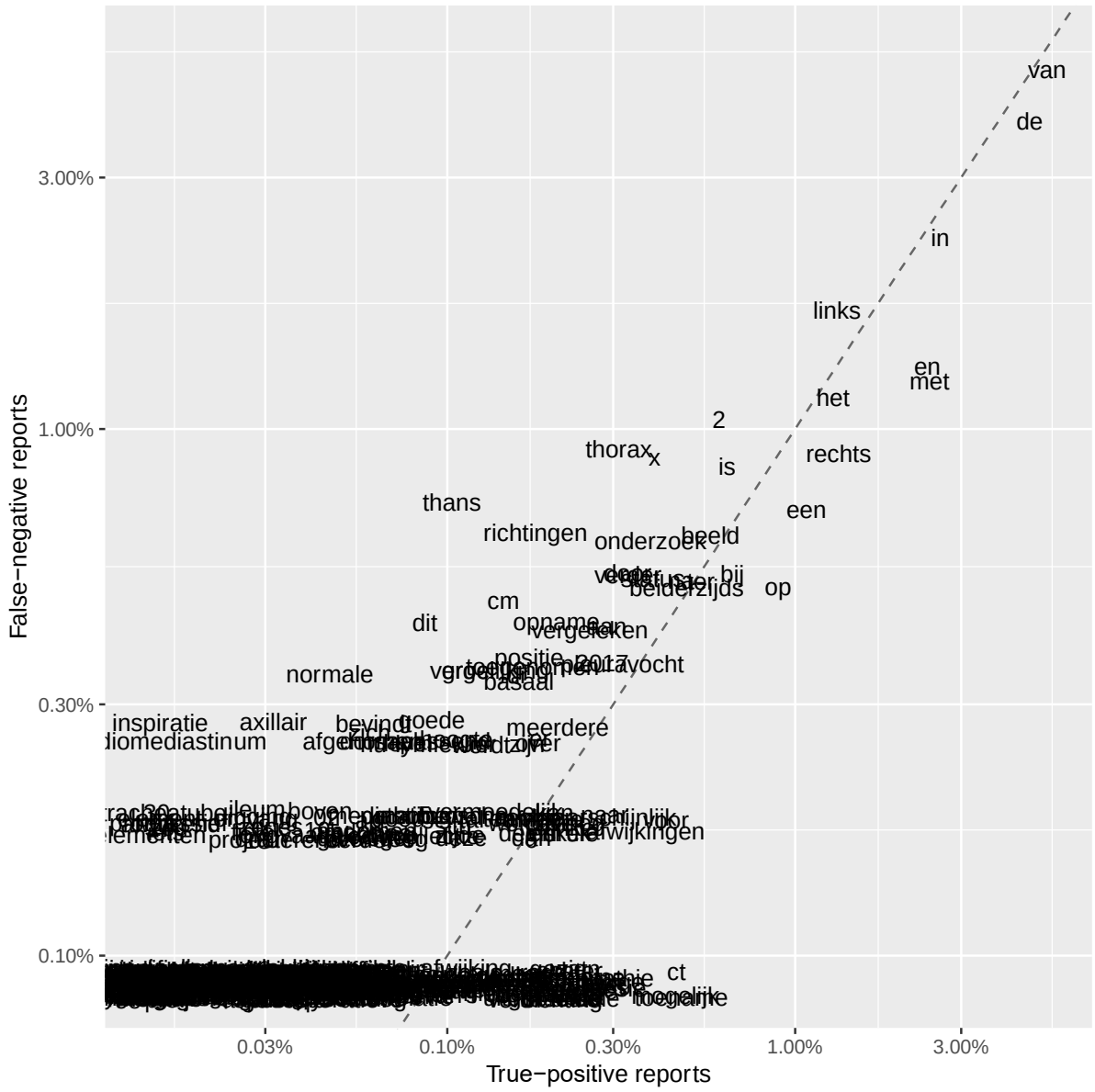

Figure S1. Scatter plot of relative word frequencies in the false-negatively and true-positively classified radiology reports. The dashed line indicates equal relative frequencies in both report sets.

Figure S2 shows the relative frequencies of words in the false-positive and true-negative reports. In the false-positive reports, words such as “mogelijk” (“possible”), “fractuur” (“fracture”), and “laesie” (“lesion”) are overrepresented. On the other hand, “normale” and “normaal” (both translations of “normal”) are overrepresented in the true-negative reports (i.e., underrepresented in the false-positive reports). Both differences may affect the classifier’s performance, possibly increasing the probability of an actionable finding in the false-positive report set.

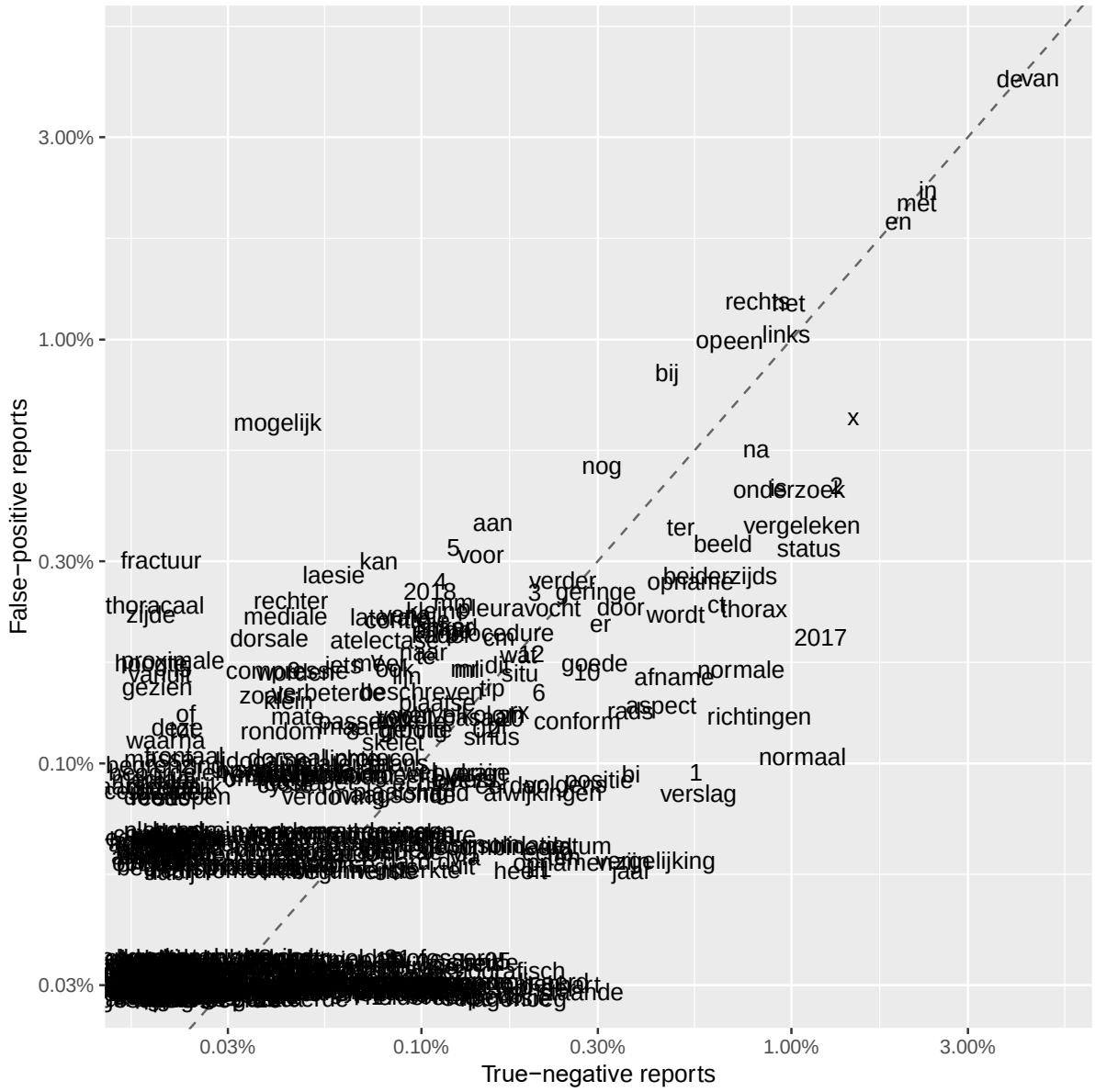

Figure S2. Scatter plot of relative word frequencies in the false-positively and true-negatively classified radiology reports. The dashed line indicates equal relative frequencies in both report sets.