Equivalence of pathologists' and rule-based parser's annotations of Dutch pathology reports

Burger, Gerard TN; Abu-Hanna, Ameen; de Keizer, Nicolette F.; Burger, Huibert; Cornet, Ronald

Published in:
Intelligence-Based Medicine

DOI:
10.1016/j.ibmed.2022.100083

Intelligence-Based Medicine 7 (2023) 100083

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Equivalence of pathologists' and rule-based parser's annotations of Dutch pathology reports

Gerard TN. Burger [a,b,*], Ameen Abu-Hanna [b], Nicolette F. de Keizer [b], Huibert Burger [c], Ronald Cornet [b]

[a] Symbiant Pathology Expert Centre, Hoorn, the Netherlands
[b] Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands
[c] Department of General Practice and Elderly Care Medicine, University, Medical Center Groningen, University of Groningen, Groningen, the Netherlands

ARTICLE INFO

Keywords:
Natural language processing
Systematized nomenclature of medicine
Automatic annotation
Pathology
Information storage and retrieval

ABSTRACT

Introduction: In the Netherlands, pathology reports are annotated using a nationwide pathology network (PALGA) thesaurus. Annotations must address topography, procedure, and diagnosis.
The Pathology Report Annotation Module (PRAM) can be used to annotate the report conclusion with PALGA-compliant code series. The equivalence of these generated annotations to manual annotations is unknown. We assess the equivalence of annotations by authoring pathologists, pathologists participating in this study, and PRAM.
Methods: New annotations were created for one thousand histopathology reports by the PRAM and a pathologist panel. We calculated dissimilarity of annotations using a semantic distance measure, Minimal Transition Cost (MTC). In absence of a gold standard, we compared dissimilarity scores having one common annotator. The resulting comparisons yielded a measure for the coding dissimilarity between PRAM, the pathologist panel and the authoring pathologist. To compare the comprehensiveness of the coding methods, we assessed number and length of the annotations.
Results: Eight of the twelve comparisons of dissimilarity scores were significantly equivalent. Non-equivalent score pairs involved dissimilarity between the code series by the original pathologist and the panel pathologists. Coding dissimilarity was lowest for procedures, highest for diagnoses: MTC overall = 0.30, topographies = 0.22, procedures = 0.13, diagnoses = 0.33.
Both number and length of annotations per report increased with report conclusion length, mostly in PRAM-annotated conclusions: conclusion length ranging from 2 to 373 words, number of annotations ranged from 1 to 10 for pathologists, 1–19 for PRAM, annotation length ranged from 3 to 43 codes for pathologists, 4–123 for PRAM.
Conclusions: We measured annotation similarity among PRAM, authoring pathologists and panel pathologists. Annotating by PRAM, the panel pathologists and to a lesser extent by the authoring pathologist was equivalent. Therefore, the use of annotations by PRAM in a practical setting is justified. PRAM annotations are equivalent to study-setting annotations, and more comprehensive than routine coding. Further research on annotation quality is needed.

1. Introduction

Pathology reports provide important diagnostic information in various settings, like in oncology. Oncological pathology is important for establishing an accurate prognosis and selecting the most appropriate therapy for the individual patient [1]. As a need exists for annotating the data, that is labeling its contents, for efficient retrieval of information for the purpose of clinical medicine and research, widespread efforts have been made to provide these [2–5]. In the field of pathology, pathologists provided their findings traditionally in free text. There are three main methods of generating annotated data [1,6–9]:

* Corresponding author. Pathologie Symbiant BV, P/a Dijklander Ziekenhuis, Maelsonstraat 3, 1624NP, Hoorn, the Netherlands.
E-mail address: g.t.burger@amsterdamumc.nl (G.TN. Burger).

https://doi.org/10.1016/j.ibmed.2022.100083
Received 11 February 2022; Received in revised form 23 October 2022; Accepted 7 December 2022
Available online 11 December 2022

1. Manual entry of codes for at least location and diagnosis but often also codes for operation and other characteristics of the lesion being reported on;
2. Synoptic reporting e.g. Refs. [10,11], from which these codes are generated;
3. Automated coding by natural language processing of free text [12–14].

The multiple codes are added to and stored with the pathology report and form one or more code series. The authoring pathologist is responsible for the correctness of the codes. The second method, synoptic reporting is mainly used for the most common types of oncological or precancerous specimens. Synoptic reporting involves capturing coded data in structured forms. The importance of the availability of annotated data is paramount in oncology. In the Netherlands, synoptic reporting by data entry using protocols implemented as computer forms is now the recommended practice in oncology, especially for breast and colon tumors [11,15–17]. For other specimens, such as less frequently occurring oncological material and benign specimens, the pathologist has to create a free-text report including appropriate codes. See Fig. 1 for an overview of the reporting process.

Whether manually entered or automatically generated from free-text or structured reports, the resulting codes need to be of sufficient quality, i.e., need to be complete and correct for clinical use and for data reuse purposes such as quality assurance or research [13,14,18]. Human authoring of codes is heuristic, with potentially limited reproducibility and inter-coder agreement. It is important to realize that in the Netherlands no consensus exists among pathologists on the best way of coding a report conclusion text, which precludes determining a coding gold standard. On the other hand, an automatic annotation system will consistently provide the same annotation, whether correct or incorrect, for a given report's conclusion text, but inter- and intra-annotator agreement among humans and an automatic annotation system is currently not known [19]. Although the PALGA database software, the Dutch national pathology report database, enforces basic coding rules as detailed below, coding practice in the field varies greatly from very concise to very elaborate annotation, a problem also known to exist in France [14].

In the Netherlands an automatic annotating system Pathology Report Annotation Module (PRAM) exists, a rule-based natural language parser for the Dutch language that takes advantage of the specialized use of the language in pathology report conclusions. The rule-based algorithms used by the PRAM are based on the expertise of the pathologist author and can be summarized as a form of heuristic natural language processing. This system, like comparable systems for other languages, may fail to recognize important linguistic constructions like ambiguous words and sentences, and misinterpret punctuation [8]. Also, negation constructs can cause problems [9]. Therefore, besides introducing the aforementioned annotating system, in this paper we research the feasibility of using this system in a practical setting by formally assessing the equivalence of the codes generated by the PRAM to those created manually by pathologists, both in clinical practice and in an experimental setting. This assessment is performed by comparing the annotations authored by the PRAM and the human pathologists for a test set of 1000 data points using a distance metric. Additionally, we compare the lengths of the resulting sequences of codes (code series), as these may indicate the amount of information represented in the codes.

2. Background (693 words)

2.1. PALGA and the PALGA thesaurus

1. In the Netherlands, the pathology reports of all pathology laboratories are stored as free text, or as structured data if synoptic reporting is used, in a distributed database system primarily for the purpose of patient care. This system is administered by the PALGA Foundation [20], established in 1971, which also facilitates scientific research using the data stored in this system. The multi-axial PALGA thesaurus is used for providing medical codes stored with the report. It is a dictionary in which PALGA codes are provided with multiple synonymous descriptions and with mappings to corresponding SNOMED CT codes. These PALGA codes were derived from SNOMED II, and are assumed to represent the findings in the conclusion section as accurately as possible. The thesaurus contains the axes T = topography, P = procedure, M (for morphology) = diagnosis, D = disease, E = etiology, F = function, and J = occupation [21]. A single PALGA code is compiled from the axis' code prefix followed by five duodecimal numbers (using digits X for decimal 10 and Y for decimal 11). Especially in the topography axis these numbers express a form of hierarchy. This hierarchy is restricted, as it is represented by the identifier, which implies limitations in depth and breadth of the hierarchy, and the impossibility of multiple hierarchies [22].

In the PALGA thesaurus, composite codes compiled from 2 or more single codes also occur. These mostly denote a topography combined with a procedure or a diagnosis, like "T62000M20400" denoting atresia (M20400) of the esophagus (T62000). These single or composite codes are then linked together with asterisks (*) to form a PALGA code series. Separate findings in one report may be annotated into one or more separate code series. The construction of these series is bound to rules [23]. The most important rule is the code sequence in the series: at least one topography has to be annotated first, followed by procedure(s), followed by one or more diagnoses. After these, arbitrary codes may be added in any sequence. For example, the conclusion text (in Dutch) "huidbiopt linker hand met lichen planus" (left hand skin biopsy with lichen planus) corresponds to this PALGA code series: "T01000*TY8700*TYY990*P11400*M48890" (i.e., skin*hand*left*biopsy*lichen planus). In clinical practice the descriptions for the PALGA codes are used: a pathologist would write or dictate the asterisk-separated line of terms "skin*hand*left*biopsy*lichen planus", and the PALGA database management system would convert these concatenated descriptive terms into the PALGA code series mentioned above.

Pathologists are required to add appropriate codes to the report, at least in one code series. Specification of codes for the axes topography, procedure and diagnosis is mandatory in the Netherlands [23]. The use of the other axes is not mandatory; hence these are less frequently used. This is also the case in for example Denmark [7] (mandatory codes for topography and diagnosis) and France [14] (mandatory codes for topography, technique, procedure and diagnosis).

2.2. PRAM

The PRAM software uses a rule-based approach to identify terms that can be assigned a PALGA code from conclusion sections of free-text pathology reports. This relieves pathologists from the duty of providing the codes, a tedious and possibly error-prone routine that might produce non-uniform results [24–27]. The system has been introduced in 2013. It adheres to the PALGA coding rules detailed above. Other specific rules for coding reports on cervical and endometrial topographies on behalf of the Dutch cervical cancer population survey apply; these are handled by the PRAM as well.

The software has a modular structure as detailed in Fig. 2.

Explanation: in the preprocessing stage, apart from stemming and converting words pathologists use to describe a certain medical entity but that are not present in the PALGA thesaurus to an equivalent term that is present, deprecated terms from the text are converted to equivalent not-deprecated terms. As stated, also numbers, dates and report identifiers are recognized for research purposes, but these data are not used in this study. In the feature extraction stage, the system performs the following steps [1]: recognize the code descriptions from the preprocessed words individually or in combination with other words in the vicinity, non-recognized words are kept [2], extract code descriptions

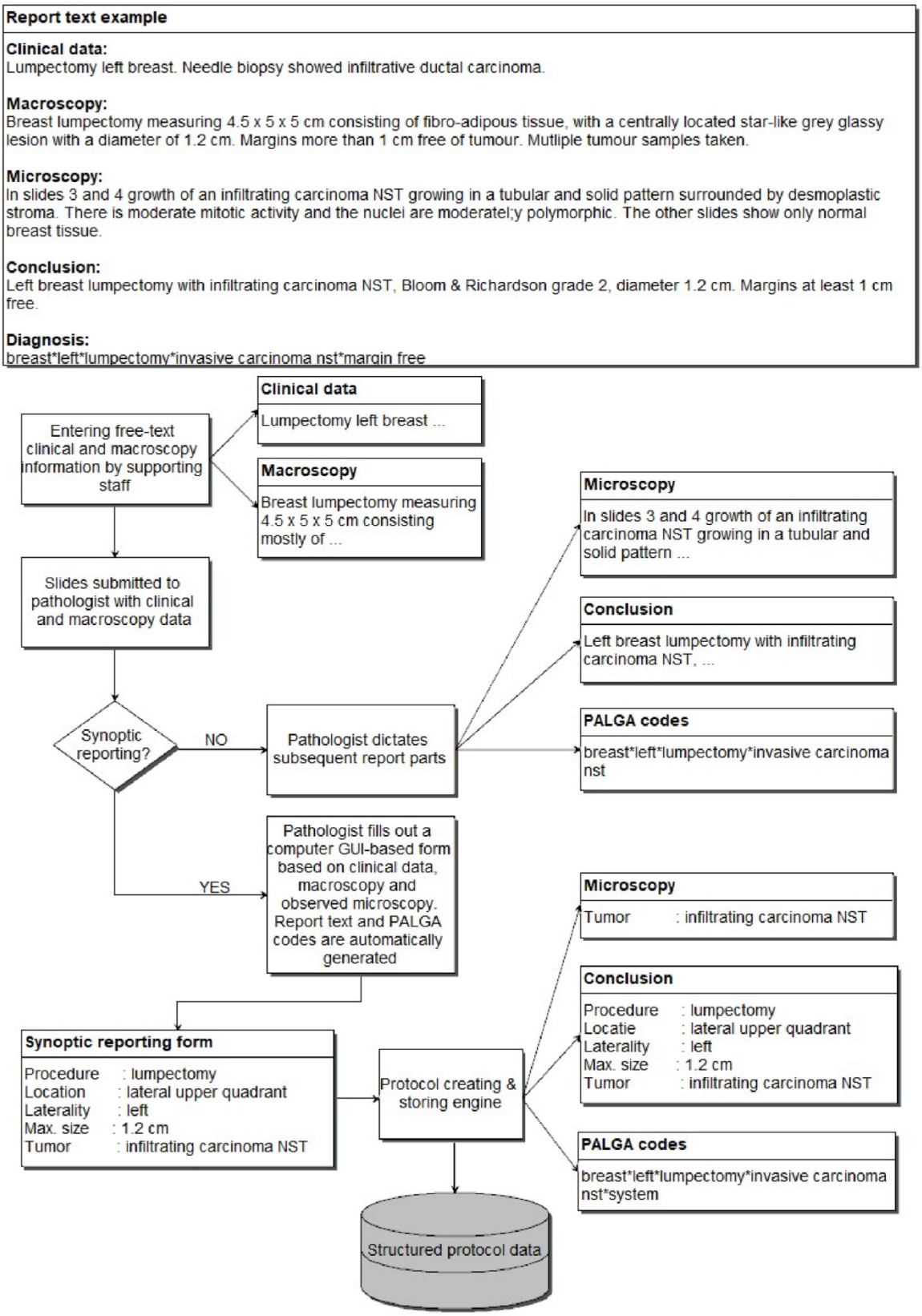

Fig. 1. Pathologist reporting process (exemplified by a lumpectomy of a carcinoma of the left breast, for PALGA code structure see Background).

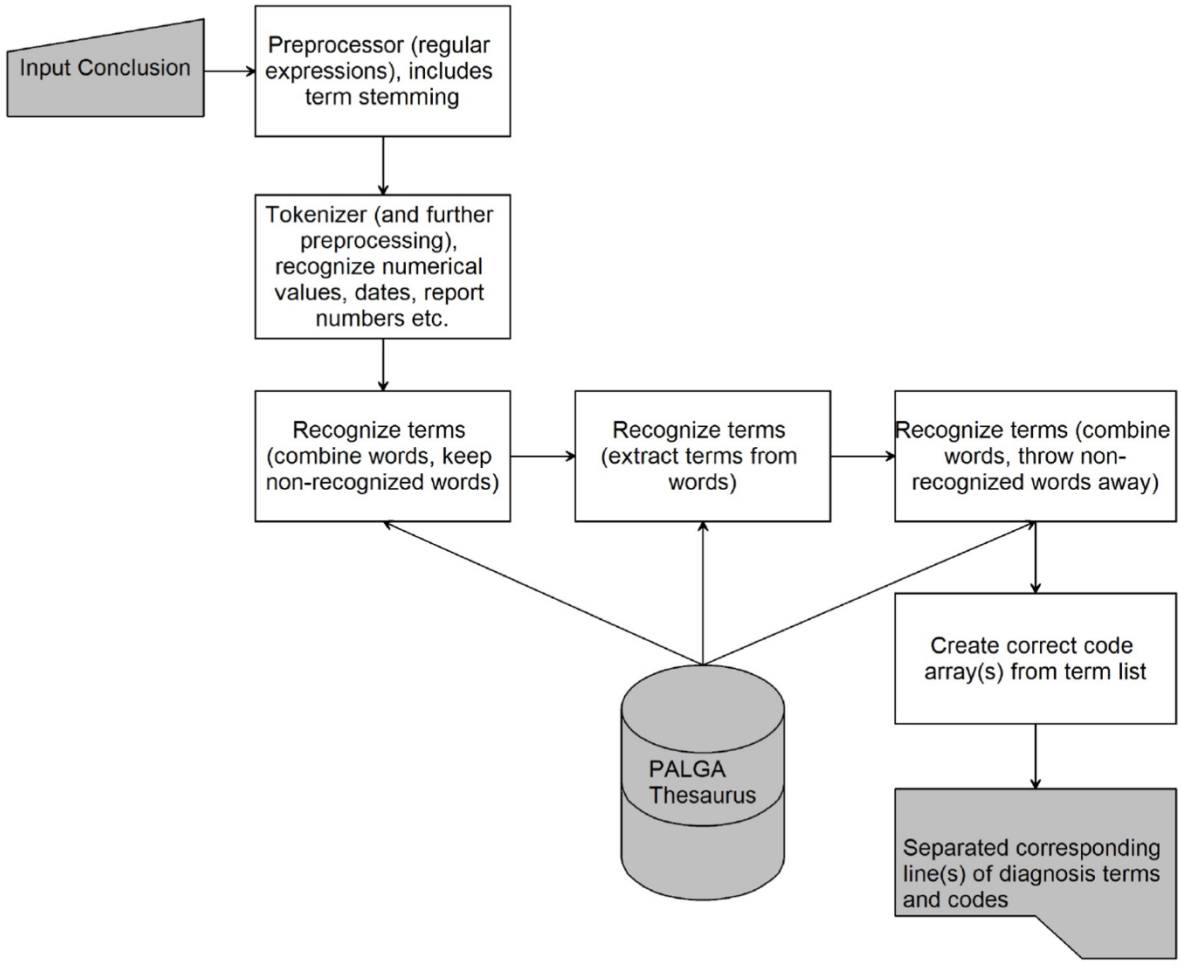

Fig. 2. PRAM workflow.

from the words and [3] repeat step 1 to discover new code descriptions by recombining the words generated in the previous steps. All non-recognized words from the text are discarded at this stage.

The PRAM has been developed in C# by the first author (GB), using the Microsoft. NET Framework, and runs under version 3.5 or higher of this framework, or under the corresponding Mono versions. Running in a Visual Studio 2022 debugging session its memory usage is 22 Mb and its CPU usage is 10% (97% kernel activity). Using a release build, the 1000 conclusions used in this study take 11 s, i.e., 11 ms for one conclusion. This was tested under Windows 10 on an Intel Core i5 processor running at 3.3 GHz. The main library function accepts a conclusion text as its input and generates PALGA terms and corresponding codes as its output.

The PRAM will generate code series based on PALGA codes that correspond to recognized entities as described in the background. As in the Netherlands free-text reports are generated by dictating into speech-recognition software typos are not an issue in this setting, although word recognition problems could lead to wrong codes. However, pathologists would recognize this as they carefully scrutinize the conclusion text from which the codes are derived. It can therefore be assumed that this does not present a major problem. The software needs at least one recognizable topography term and one recognizable diagnosis term to generate a code series.

3. Materials & methods

The PALGA database contains over 67 million annotated pathology reports from 1971 until now. From these, as a test set we selected from PALGA a random (regarding specimen, pathologist, and institute) sample of 1000 conclusions and PALGA code series from histology reports from 2015 that had been manually coded. The PRAM does not use a machine learning approach, therefore no need existed to create a training or validation data set.

To keep the workload reasonable for the participating pathologists, the sample size was limited to 1000. Cytology, autopsy or molecular pathology reports were not included in the sample.

We used histology reports alone because cytology report conclusions are in general very simple. Therefore, we felt that comparison of annotation by the PRAM and by pathologists would not yield results of any significancy. Moreover, the PRAM is used much more frequently for annotating histology reports than for annotating cytology reports. On the other hand, autopsy conclusion texts tend to be relatively large and complex, and such texts are not handled well by the PRAM, which relies on the simplified Dutch grammar used by pathologists on the relatively short conclusions in histology reports. Molecular pathology reporting uses a quite different style of reporting which is not covered by the PRAM.

The 1000 test set conclusion texts were coded by the PRAM software. Four volunteering pathologists from the pathology laboratory in which one of the authors (GB) is employed each annotated 500 conclusion texts: pathologist A annotated conclusions 1–500 inclusive, pathologist B 251–750, pathologist C 501–1000, pathologist D 1–250 and 751–1000. Thus, each report conclusion was annotated by two pathologists. Three of the pathologists had more than five years of experience in pathology practice, one (A) had approximately one year of experience. The pathologist author of this paper, GB, did not participate in the coding. To create a coding “silver standard”, pathologists were

instructed to annotate the conclusions as completely as possible using the descriptions of PALGA codes (as they do in daily practice). No limits were set on the use of codes or on the number of code series per report. The original code series and PRAM-generated code series were not available to the study pathologists.

In the four manually-created sets of descriptions for the codes in the code series artifacts like spacing or erroneous hyphens were removed by one of the authors (GB), for all other errors, the coding pathologist was asked to re-code the conclusion. Any perceived semantic errors or coding errors were not corrected.

Four sets are mutually compared, using the method described below: the original set of codes series by the original authoring pathologist (AP), second (pathologist 1 participating in the study, PP1) and third (pathologist 2 participating in the study, PP2), fourth (PRAM), the series generated by the PRAM.

3.1. Dissimilarity measure of codes

As stated above, apart from a small set of annotation rules, there is no agreement on the best way of coding pathology reports. Therefore, conventional metrics such as sensitivity and specificity are not applicable here, because a gold standard to compute these metrics against is not available. Instead, we relied on the use of a distance metric to assess dissimilarity of annotations. We used a method from natural language processing: a modified version of the Levenshtein edit distance. To determine the distance of the entities represented by the individual PALGA codes we converted these to equivalent SNOMED CT codes, of which the inter-entity distance can be easily determined because of the hierarchical structure of the SNOMED CT ontology. This process is detailed below.

To compare code series, we devised a dissimilarity measure of two

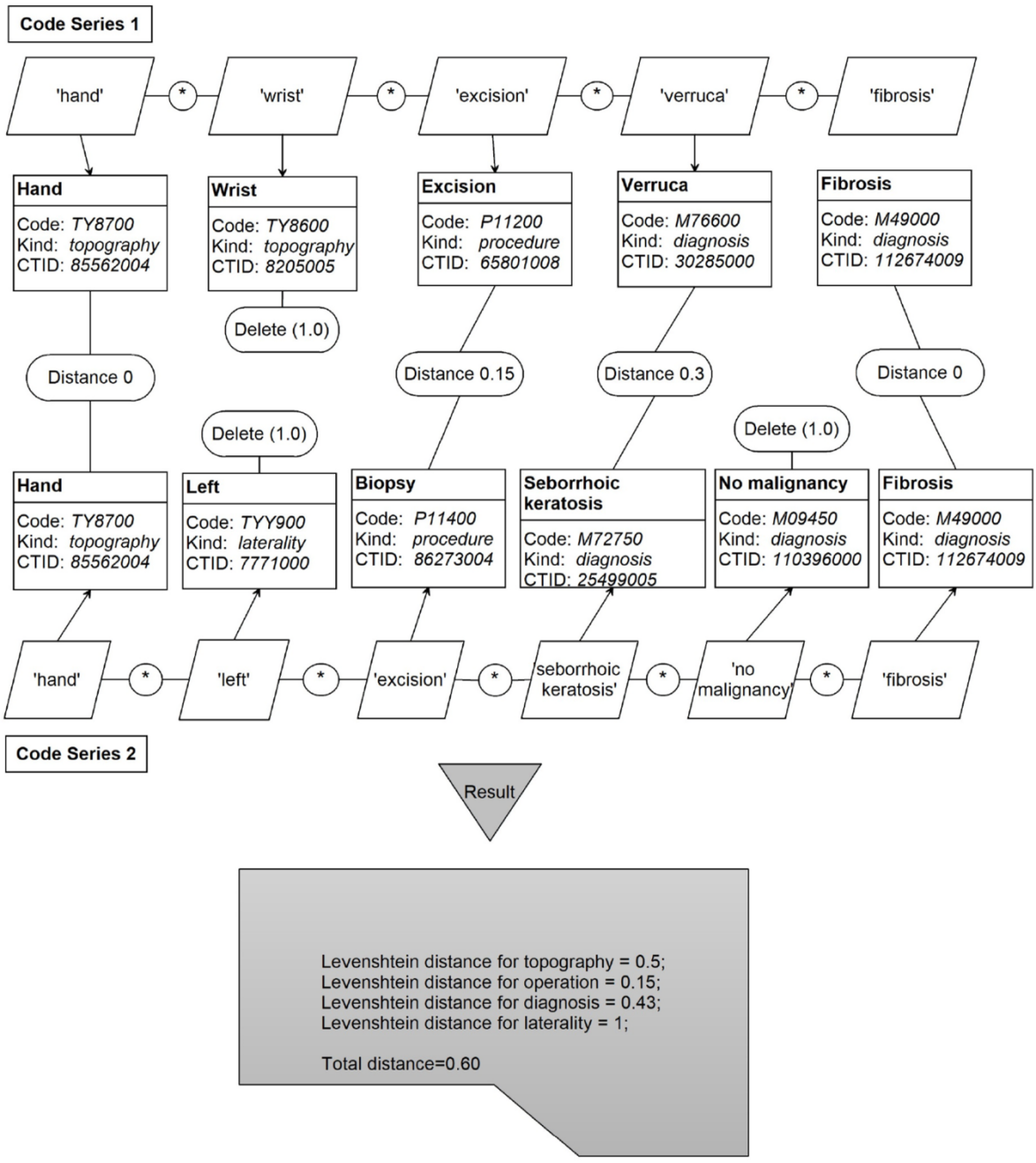

Fig. 3. Comparison of two PALGA code series. Deletion cost as well as insertion cost of one code is 1.0.

individual PALGA codes. This measure takes into account that, for example, “hand” and “wrist” are more similar topographies than “hand” and “foot”. As the limited hierarchy in PALGA codes is not sufficient for this [28] we applied the is-a hierarchy of SNOMED CT, utilizing the SNOMED CT codes provided by the PALGA thesaurus corresponding to the used PALGA codes. We defined (normalized) inter-code distance or dissimilarity as the distance between the two codes as defined in the SNOMED CT database divided by the largest inter-code distance occurring in the SNOMED CT database. An example is displayed in Fig. 3.

3.2. Dissimilarity measure of code series

For each conclusion, code series were merged into a single series if there were more than one for the purpose of this research. Comparison between two code series involves both alignment of code series and comparison of the individual codes.

Alignment was needed to compare only the most similar codes. For this, both (lumped) code series are split into code sets for topographies, procedures, diagnoses and laterality indications. The minimal distances (minimal transition cost, MTC) between these four code sets are then determined by applying a modified version of the Levenshtein distance (MLD) [29], where code order is not taken into account. Thus, the MTC between two code series is the measure of their dissimilarity.

The computation is described in Table 1.

This yields a number between zero and one, where zero means complete similarity and 1 complete dissimilarity. An example is shown in Fig. 3.

For each conclusion, total MTC was calculated for all six combinations of code series: AP (Authoring Pathologist) vs. PP1 (Participating Pathologist 1), AP vs. PP2 (Participating Pathologist 2), AP vs. PRAM (Pathology Report Annotation Module), PP1 vs. PP2, PP1 vs. PRAM, PP2 vs. PRAM.

The equivalence of PRAM coding to human coding is measured by comparing sets of MTCs between the several coding pairs formed by the human coders and the PRAM.

We compared MTCs for all coder pairs that had one coder (AP, PP1, PP2, PRAM) in common, where order in coder pair did not matter (comparing AP-PP1 to AP-PP2 is the same as comparing AP-PP2 to AP-PP1), yielding 12 combinations.

As a supplementary analysis, we assessed, using the ‘accuracy’ classification metric, the agreement between PRAM on the one hand and pathologist panels PP1 and PP2 on the other hand. To this end, we assumed that a MTC lower than the threshold of 0.50 indicates ‘agreement’ between coders. We also measured the average code count and the average code series count (before merging) as a function of the conclusion word count as a measure of coding comprehensiveness.

Table 1
Workflow of numerical comparison of 2 (sets of) PALGA code series.

Step	Action
1	If more than one code series are in a set, lump the codes, creating one code series
2	Split composite codes into single ones
3	Remove duplicate codes
4	Split codes series into sets of codes for: topography, procedure, diagnosis and laterality. Retain code count of the largest series
5	Compute the minimal transition cost (MTC) for codes from the abovementioned sets examining all possible combinations within these sets (insertion/deletion cost are 1, replacement cost is inter-code distance as defined above) using Knuth’s non-recursive permutation generating algorithm [30]. The code count before the removal of duplicate codes is used in the computation of the MTC
6	Calculate the total minimal transition cost, normalized by division by sqrt [4]: $\mathrm{MTC}_{\mathrm{total}} = \frac{\sqrt[2]{\mathrm{MTC}_{\mathrm{topographies}}{}^2 + \mathrm{MTC}_{\mathrm{lateralities}}{}^2 + \mathrm{MTC}_{\mathrm{procedures}}{}^2 + \mathrm{MTC}_{\mathrm{diagnoses}}{}^2}}{2}$

3.3. Statistical analysis

As set out above, MTCs for coder pairs were compared for 12 combinations by calculating their differences for each of the reports and calculating their mean. The distribution of each of the 12 differences was checked for normality before the analysis of equivalence. We applied equivalence testing, i.e., we assumed as a null-hypothesis that the two groups are different (non-equivalent), on average, by a certain amount delta or more. A difference between coder pairs was regarded equivalent, being the alternative hypothesis, if its mean value and 95% confidence interval (CI) fell within a range delimited by minus delta and plus delta. Delta is the equivalence margin defined as a range of values for the differences that were regarded small enough to consider coding pairs equivalent. A meaningful equivalence margin is hard to determine and is dependent on many factors [31,32]. It should be a fraction of the expected effect (in our case, the MTC). In this study we applied the strictest delta we found in the literature [33,34], which is 0.1.

3.3.1. Accounting for clustering

In this study, multiple observations are made in a number of distinct groups, meaning that the data are clustered [35]. These groups or “clusters” were defined by the coding pathologists. Conclusion texts were annotated by the pathologists participating in the experiment as stated above. Therefore, MTC values within a cluster may be more similar than MTC values from different clusters. This phenomenon creates a correlation between MTCs within the same cluster. Ignoring substantial clustering may seriously affect the results of statistical analysis. Therefore, to assess whether accounting for clustering in the analysis was needed we calculated the “design effect” which indicates the extent to which the standard error of the estimate of the mean is underestimated in clustered samples [36]. It is equal to 1 + (average cluster size - 1) * intraclass correlation (ICC). Cluster size was the size of the coding of one pathologist’s group, i.e., 250. The ICC is a measure of correlation within the cluster (group) and equal to the proportion of the total variance accounted for by between-group variance. The ICC was estimated by invoking a so-called empty multilevel regression model which contained only an intercept, i.e., no explanatory variables, and a random effect for coder group. The difference in MTC served as the dependent variable. With the ICCs from these analyses the design effect was calculated and when it appeared larger than two [37] the mean differences and their 95% CI were calculated based on these empty multilevel regression models.

3.4. Outcome measures

In this study we consider the use of the PRAM to generate code series in a clinical setting feasible if the annotation by the PRAM is equivalent with, i.e., not statistically significant dissimilar from, the annotation by the original annotators or from the annotation by the pathologists participating in this study. To this end, we measured and compared dissimilarity (expressed as MTC) of code sets generated by different coders (the original coders, the 4 pathologists generating 2 sets of code series, and the PRAM). Additionally, we assessed and compared the length of the resulting code series, as a proxy for the quality of the codes.

4. Results

Of the 1000 reports used in this study, 965 were used to calculate the MTCs of the generated code series. Thirty-one (3%) reports could not be annotated by the PRAM and were excluded from further processing. This exclusion is justified because all of these reports had a missing topography and/or diagnosis in the conclusion (17 missing topographies and 23 missing diagnoses in 31 reports), indicating that the report is not acceptable and hence preventing coding by the PRAM. Four reports were not annotated by one of the participating pathologists, without any given reason, these were excluded as well. In the codes of the authoring

pathologist (AP) no errors were found. In the codes used in the set created by pathologist PP1, 82 (13%) non-existent codes were found, 87 (6%) code series were corrected. In the codes used in the set created by pathologist PP2, 127 (19%) non-existent codes were found, 196 (14%) code series were corrected.

The average MTCs are given in Table 2 (overall distance and distances within the different axes in the PALGA thesaurus are given).

The outcomes of the statistical analyses of the comparison of series of overall MTCs are shown in Table 3.

For the majority (N = 8) of comparisons, the design effect was larger than two and consequently the multilevel regression approach to calculating the mean difference in agreement scores was followed for all comparisons.

Table 3 shows that:

1. Comparisons of MTCs involving AP show statistically significant equivalence between the results of PP1, PP2 and PRAM;
2. Comparisons of MTCs where PP1 participates in both comparisons show statistically significant equivalence between PP2 and PRAM, but no statistically significant equivalence between PP2 and AP; likewise, comparisons of MTCs where PP1 participates in both show statistically significant equivalence between PP1 and PRAM, but no statistically significant equivalence between PP1 and AP;
3. Comparisons of MTCs where PRAM participates in both show statistically significant equivalence between PP1 and PP2, but no statistically significant equivalence between PP1 and AP, and PP2 and AP.

The accuracy of coder pairs PRAM-PP1 and PRAM-PP2 (threshold MTC = 0.5) was as follows: PRAM-PP1 = 89%, PRAM-PP2 = 88%.

The count of codes used by the coders compared to the conclusion word count is shown in Fig. 4.

Correlation coefficient for AP 0.64 (moderate), PP1 0.64 (moderate), PP2 0.54 (moderate), PRAM 0.85 (high).

The count of code series generated by the coders compared to the conclusion word count is shown in Fig. 5.

Correlation coefficient for AP 0.62 (moderate), PP1 0.61 (moderate), PP2 0.53 (moderate), PRAM 0.82 (high).

5. Discussion

In this study we introduced a rule-based Dutch natural language parser, and we aimed to determine the equivalence of annotation by this rule-based natural language parser and by pathologists coding Dutch pathology conclusions, using MTC as a measure of dissimilarity. Our results showed that the MTCs of PP1-to-PP2, were equivalent to those of PP1-to-PRAM and PP2-to-PRAM. Also, the MTCs of AP-to-PP1, AP-to-PP2 and AP-to-PRAM were equivalent. The MTCs of AP-to-PP1 and AP-to-PP2 on the one hand, and PP1-to-PP2 on the other hand were not demonstrated to be equivalent, and also the MTCs of AP-to-PRAM and both PRAM-to-PP1 and PRAM-to-PP2 did not show equivalence. This is visualized in Fig. 6.

Table 2
Mean dissimilarity measures. PP1 = participating pathologist 1; PP2 = participating pathologist 2; AP = authoring pathologist; PRAM = Pathology Report Annotation Module.

	MTC overall	MTC topography	MTC procedure	MTC diagnosis	MTC laterality
AP vs. PP1	0.31	0.19	0.11	0.34	0.18
AP vs. PP2	0.31	0.19	0.12	0.33	0.19
PP1 vs. PP2	0.23	0.15	0.10	0.26	0.07
AP vs. PRAM	0.36	0.27	0.13	0.42	0.19
PP1 vs. PRAM	0.27	0.25	0.14	0.30	0.05
PP2 vs. PRAM	0.29	0.24	0.15	0.32	0.06
Average	0.30	0.22	0.13	0.33	0.12

The accuracy results of PP1-to-PRAM and PP2-to-PRAM support the equivalence of annotation by the PRAM on one hand, and of annotation by the pathologist panels PP1 and PP2 participating in this study, when PP1 and PP2 are considered to be the gold-standard. We therefore conclude that the performance of a rule-based parser is equivalent with the performance of pathologists who are expressly asked to annotate the report conclusion as comprehensively as possible (PP1 and PP2), and also, but perhaps to a lesser extent, with the performance of pathologists annotating in routine practice (AP). For each pair-wise comparison the diagnosis axis exhibited the highest dissimilarity in coding, and the procedure and laterality axes the smallest dissimilarities. This could partly be explained by the number of codes in the PALGA thesaurus: 7505 diagnosis codes, 2334 topography codes, 388 procedure codes, and 15 laterality codes in the version of the thesaurus used in this study. Also, a pathologist will spend most effort to precisely describe the diagnosis, this being the most significant part of the report conclusion.

Annotating pathology reports in a clinical setting is no common practice, hence the literature on comparing manual with automatic coding is sparse. As the available studies all take a different approach in comparing automatic annotation with manual coding, it is not possible to directly compare results. An earlier Dutch study [12] adopted a nearest neighbor approach, suggesting several coding alternatives for a given report. In this study the authors adopted a comparison method between PALGA codes based on the crude hierarchy present in SNOMED II [28]. A French study [14] also presented several coding alternatives for one report using weighing of word sequences for the classification of a report. However, it did not state how the generated alternatives were judged against the original codes. Of note, consensus on the topography was reported to be lowest in this French study, whereas in our study most dissimilarity is found in the diagnosis axis. The reason might be a much larger topography/diagnosis proportion in the PALGA thesaurus as compared to the ADICAP codification system used in the other study, and the presence of several topography codes in the conclusion. A second French study [13] took a supervised machine-learning approach. Again, the comparison method for the generated annotations with the pre-existing manual codes was not stated. When classifying reports using ICD-O3, again topography was the most difficult to code completely, although this partially may have been due to lacking information on this axis in the examined reports. The number of reports used in these studies varied strongly from 640 to 15,000 reports. All these studies used the whole report text instead of the conclusion only, as we did. Using the whole report text may be advantageous for the granularity of the annotation, as suggested by Ref. [12]. None of these studies made use of re-annotating experts, which we consider an added value of our study, enabling comparison of two modes of annotation: in daily practice, and with (presumably) extra attention for the annotation process.

This study supports the hypothesis that it is possible to create automatic annotation for Dutch pathology reports that is comparable to annotations created by pathologists who were explicitly asked to create comprehensive codes, and with codes created in routine practice. There are several other studies using parser methods for natural language processing (NLP) of pathology reports in different languages [2,14,38–40], suggesting that this method is feasible for use in other languages than Dutch. The annotation generated by the PRAM is more elaborate than the original codes, and therefore likely to be more representative for the corresponding conclusion text. The performance of a rule-based parser depends on the quality of the input, e.g., the absence of typing errors, which is the case in present Dutch pathology practice where dictating is common practice. If such errors could occur, another pre-processing module correcting these errors before further handling of the text would have to be added, or one could resort to neural NLP

Table 3
Comparison of MTCs. PP1 = participating pathologist 1; PP2 = participating pathologist 2; AP = authoring pathologist; PRAM = Pathology Report Annotation Module. The numbers are the mean of the MTC differences between coder groups (95% CI). Green background indicates equivalence of the dissimilarity scores (MTC's), i.e., their mean differences and 95% CIs are within a range delimited by minus 0.1 and plus 0.1, orange background no evidence of equivalence.

	AP -PP1	AP -PP2	PRAM-PP1	PRAM-PP2	PP1 -PP2
AP -PRAM	0.052 (0.037; 0.066)	0.049 (0.038; 0.060)	0.091 (0.074; 0.108)	0.077 (0.054; 0.101)	
AP -PP1		−0.003 (−0.028; 0.022)	0.039 (0.009; 0.070		0.082 (0.064; 0.101)
AP -PP2				0.028 (0.005; 0.051)	0.085 (0.063; 0.107)
PRAM-PP1				−0.014 (−0.044; 0.015)	0.043 (0.015; 0.070)
PRAM-PP2					0.057 (0.042; 0.072)

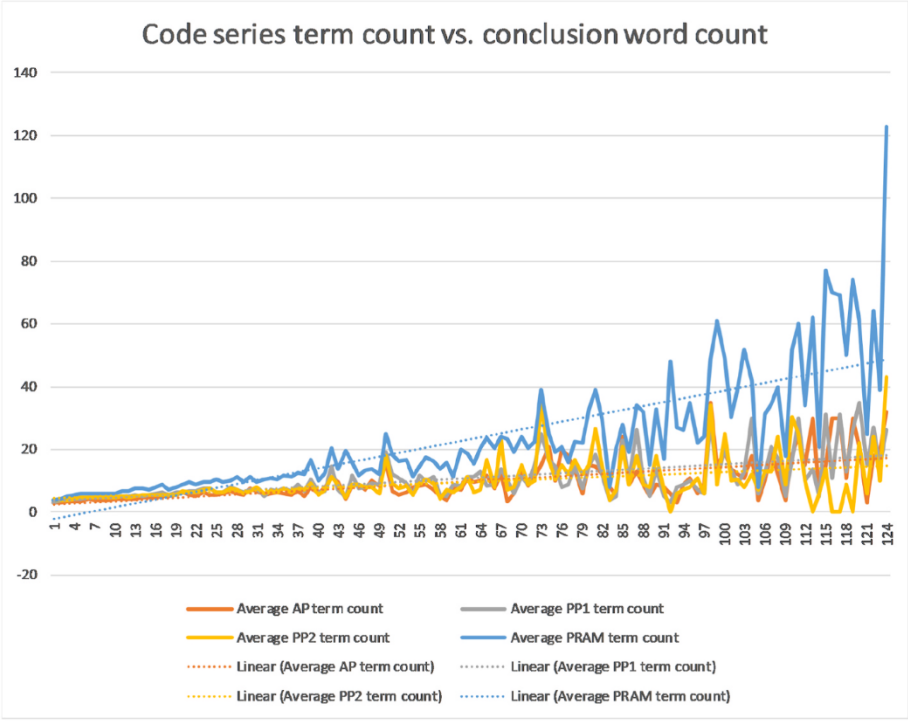

Fig. 4. Count of used codes vs. conclusion word count. X-axis = conclusion word count, y-axis = code series term count. The dashed line shows the linear trend line.

approaches such as word embeddings (contextualized or otherwise) where related words and typos would be embedded as vectors in the same neighborhood.

The study enables determining the validity of the PRAM for coding in the sense that the codes it produces are comparable to the pathologists in PP1 and PP2. The method used to calculate the similarity between two codes in a semantic network like SNOMED CT (derived from SNOMED II-like PALGA codes) is comparable to the method used by Rada et al. [41] This method has the advantage of simplicity, while the disadvantage of not taking multiple inheritance [42] does not apply in a simple context of topography and diagnosis hierarchies as is used here. Applying the modified Levenshtein method then allows us to compute a quantitative minimal transition cost. Combining these two methods yields a useful method to compute MTCs between series of codes that are connected by a simple is-a relation. This method of pair-wise comparison based on MTCs might be useful in other cases where a hierarchical coding system is used, and where a gold standard is lacking. The validity of the results is enhanced by applying the strictest equivalence margin we could find in the literature in the statistical analysis. Given our interest in patterns of equivalence and nonequivalence rather than in demonstrating equivalence of ratings per se, we refrained from controlling the "overall significance level", i.e., we did not adjust for multiple comparisons.

Some limitations of the study need to be discussed. Possible speech recognition errors may impact the performance of the PRAM more than the performance of pathologists, who understand what should have been in the conclusion text when coding. In the selection of reports for this study, reports created using the PALGA synoptic reporting software were excluded; therefore, there might be a negative selection bias concerning oncological resections of tumors of mamma, colorectal and urogenital system. We tried to minimize this effect by choosing a less recent dataset as use of synoptic reporting is increasing over time in the Netherlands, while annotation guidelines for pathologists have not changed since 2015 [23].

The question is whether the annotations generated by the PRAM actually are a better representation of the report content than the manual annotation. However, this is hard to answer. Quantitatively, human coders demonstrate smaller increase of used codes count and of code series count with increasing conclusion text word count compared to the PRAM. This might indicate higher comprehensiveness in coding by the PRAM, but this would also require qualitative assessment of code quality, which cannot be deduced from the results of this study. Because of the lack of consensus among Dutch pathologists about correct coding of report conclusions, it is not possible to establish a real gold coding standard, and hence we feel that it is also not feasible to design a study to determine if one way of coding is better than another, either produced by a human annotator or by an automatic annotating system.

The quality of the search facilities in the PALGA database depends for

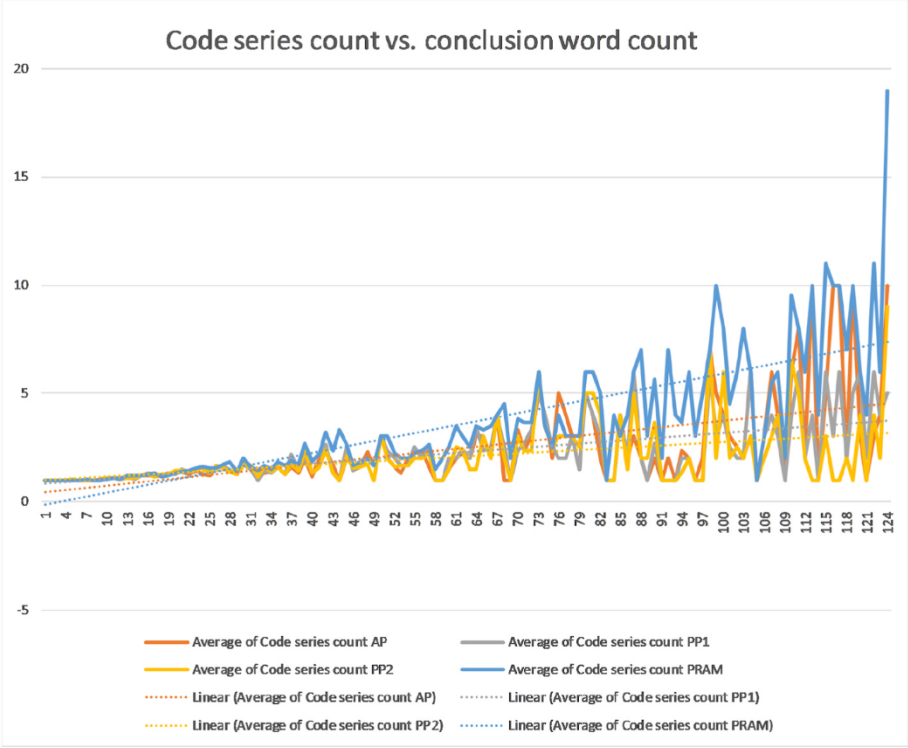

Fig. 5. Count of used code series vs. conclusion word count. X-axis = conclusion word count, y-axis = code series count. The dashed line shows the linear trend line.

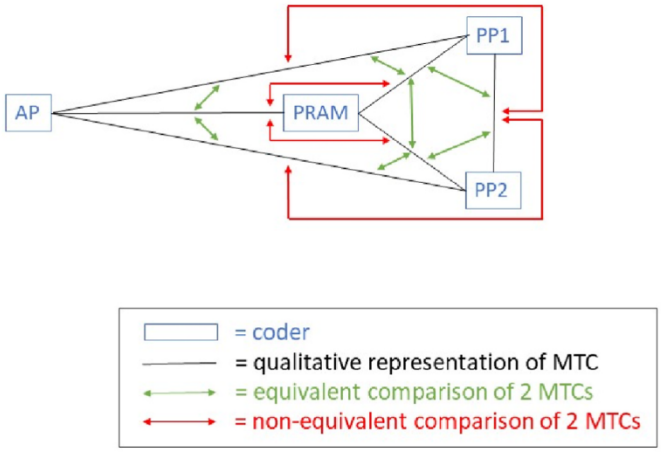

Fig. 6. Graphical representation of comparison of agreement of coder pairs (not to scale).

a great deal on the quality and quantity of the PALGA codes supplied with the pathology reports submitted to the PALGA database. This is important, because these codes are used in clinical practice providing the pathological history of a patient as well as in scientific research, providing access to the biobank formed by the pathology laboratories in the Netherlands. The annotating system introduced in this paper provides more detailed codes than human pathologists, and the significance of the PRAM might be its ability to consistently generate more codes than pathologists in clinical practice do, thus providing higher sensitivity for queries on the PALGA database. In doing this, it might outperform systems relying on machine learning: these might be made to learn to code like a human pathologist, but never to produce more extensive annotations than a pathologist would do, as these systems would have to be trained on datasets from the PALGA database, which is very large, but almost completely annotated by human pathologists.

In clinical practice, we find that while processing large and complex conclusion texts pathologists provide better annotations than the PRAM, while it seems to be the other way around with conclusion texts consisting of one of two sentences, although this cannot be concluded from the results of this study. The combination of heuristic text parsing by the PRAM for small conclusions and a machine learning system for larger conclusions might be the most consistent way of automatically generating annotations for all forms of pathology conclusion texts. This is currently under investigation. Also, further research is warranted to determine whether a rule-based parser such as the one introduced in our study can be used for reliable retrieval of clinically useful parameters such as tumor kind, molecular data and lymph node status regarding metastatic tumor when no synoptic reporting has been used. This will be the subject of future research, as is potential use of this software as a critiquing system, i.e., one that suggests relevant codes to a coding pathologist.

6. Conclusion

It is possible to reliably produce annotations of a free-text conclusion of a pathology report which are equivalent with those produced by a pathologist in routine care using a rule-based natural language parser, in this case the PRAM introduced here. This study justifies the use of such a system in a clinical setting. However, more research on the quality of the generated annotation and for a further assessment of the usefulness of this system is needed to relieve the report author from this chore, and create annotation with a possibly more constant quality. Also, evaluation of the use of a rule-based parser as a tool for retrieval of clinically relevant data is warranted.

Availability of data and materials

The datasets used in this study are available on request from the corresponding author.

Ethical approval

No patient-identifying parts of the pathology reports used in this paper were used or known to the authors. Therefore, no ethical approval was requested.

Funding

No funding was received for this study.

Author's contributions

GB developed the PRAM software, co-designed the research protocol, performed the analytical calculations and authored the manuscript. AA and NK participated in discussions on the research methodology and interpretation of the results and provided feedback on previous versions of the manuscript. HB performed the statistical analyses on the results. RC co-designed the research protocol, critiqued and edited the manuscript and was general supervisor of this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like the pathologists participating in the annotation experiment: MJ Flens MD PhD, N Gilhuijs MD, MJ de Rooij MD and M Visser MD.

References

[1] Srigley JR, McGowan T, Maclean A, Raby M, Ross J, Kramer S, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol 2009;99(8):517–24.
[2] Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016;69(11):949–55.
[3] Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inf Decis Making 2008;8(Suppl 1):S2.
[4] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inf 2009;42(5):760–72.
[5] Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:270–4.
[6] Ou Y, Patrick J. Automatic population of structured reports from narrative pathology reports. Australasian Workshop on Health Informatics and Knowledge Management 2014:41–50.
[7] Bjerregaard B, Larsen OB. The Danish pathology register. Scand J Publ Health 2011;39:72–4.
[8] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inf 2009;42(5):937–49.
[9] Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg 2007;205(5):690–7.
[10] Lankshear S, Srigley J, McGowan T, Yurcan M, Sawka C. Standardized synoptic cancer pathology reports - so what and who cares? A population-based satisfaction survey of 970 pathologists, surgeons, and oncologists. Arch Pathol Lab Med 2013;137(11):1599–602.
[11] Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Arch : an international journal of pathology 2016;468(6):639–49.
[12] de Bruijn LM, Hasman A, Arends JW. Automatic SNOMED classification–a corpus-based method. Comput Methods Progr Biomed 1997;54(1–2):115–22.
[13] Jouhet V, Defossez G, Burgun A, le Beux P, Levillain P, Ingrand P, Claveau V. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med 2012;51 (0026-1270 (Print)).
[14] Happe A, Cuggia M, Turlin B, Le Beux P. Design of an automatic coding algorithm for a multi-axial classification in pathology. Stud Health Technol Inf 2008;136:815–20.
[15] Kuijpers CCHJ. Quality assessment and improvements in pathology practice. University Medical Centre Utrecht; 2016.
[16] Snoek A, Hugen N, Visser O, Overbeek LI, Nagtegaal ID. The impact of standardized structured reporting of pathology reports for breast cancer in The Netherlands. Virchows Arch 2019;475:S18–9.
[17] Baranov NS, Nagtegaal ID, van Grieken NCT, Verhoeven RHA, Voorham QJM, Rosman C, et al. Synoptic reporting increases quality of upper gastrointestinal cancer pathology reports. Virchows Arch : an international journal of pathology 2019;475(2):255–9.
[18] Cernile G, Durbin EB. Automated classification of cancer pathology reports. In: NAACCR 2011 conference. Louisville Kentucky; 2011.
[19] Dalianis H. Pathology reports. Clinical Text Mining: Springer Open; 2018. p. 23.
[20] Casparie M, Tiebosch AT, Burger G, Blaauwgeers H, van de Pol A, van Krieken JH, et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cell Oncol 2007;29:19–24.
[21] CoA Pathologists. Systematized nomenclature of medicine 1987:594. second ed. Illinois.
[22] Cimino JJ. Review paper: coding systems in health care. Methods Inf Med 1996;35(4–5):273–84.
[23] Foundation P. Handleiding coderen. PALGA Foundation; 2008. Available from: http://www.palga.nl/assets/uploads/Coderen/Handleiding%20Coderen%20augustus%202008.pdf.
[24] Biese KJ, Forbach CR, Medlin RP, Platts-Mills TF, Scholer MJ, McCall B, et al. Computer-facilitated review of electronic medical records reliably identifies emergency department interventions in older adults. Acad Emerg Med 2013;20(6):621–8.
[25] Carter KJ, Rinehart S, Kessler E, Caccamo LP, Ritchey NP, Erickson BA, et al. Quality assurance in anatomic pathology: automated SNOMED coding. JAMIA 1996;(3):270–2.
[26] Shah AD, Martinez C, Hemingway H. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. BMC Med Inf Decis Making 2012;12(88):1–13.
[27] Hall PA, Lemoine NR. Comparison of manual data coding errors in two hospitals. J Clin Pathol 1986;39(6):622–6.
[28] de Bruijn LM. Chapter 5 - expert evaluation. In: Automatic classification of pathology reports (thesis). Maastricht: University of Maastricht; 1997. p. 88–112.
[29] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 1966;10(8):707–10.
[30] Knuth DE. A draft of section 7.2.1.2: generating all permutations. The art of computer programming, vol. 4. Addison-Wesley; 2002. p. 1–26.
[31] Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. Contr Clin Trials 2002;23:2–14.
[32] Ng T. Choice of delta in equivalence testing. Drug Inf J 2001;35:1517–27.
[33] Gao F, Ng GY, Cheung YB, Thumboo J, Pang G, Koo WH, et al. The Singaporean English and Chinese versions of the EQ-5D achieved measurement equivalence in cancer patients. J Clin Epidemiol 2009;62(2):206–13.
[34] Frey S, Dagan R, Ashur Y, Chen XQ, Ibarra J, Kollaritsch H, et al. Interference of antibody production to hepatitis B surface antigen in a combination hepatitis A/hepatitis B vaccine. J Infect Dis 1999;180:2018–22.
[35] Galbraith S, Daniel JA, Vissel B. A study of clustered data and approaches to its analysis. J Neurosci : the official journal of the Society for Neuroscience 2010;30(32):10601–8.
[36] Kish L. Survey sampling. New York: Wiley; 1965.
[37] Muthén B, Satorra A. Complex sample data in structural equation modeling. In: Marsden PV, editor. Sociological methodology. Oxford, England: Blackwell; 1995. p. 267–316.
[38] Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium 2003:584–8.
[39] Schlangen D, Stede M, Bontas EP. Feeding OWL: extracting and representing the content of pathology reports. NLPXML '04. In: Proceedings of the Workshop on NLP and XML, 4; 2004. p. 43–50.
[40] Weegar RF, Nygard J, Dalianis H. Efficient encoding of pathology reports using natural language processing. 2017. p. 778–83.
[41] Rada R, Mili H, Bichnell E, Blettner M. Development and application of a metric om semantic sets. IEEE Trans Syst Man Cybern 1989;9:17–30.
[42] Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inf 2007;40(3):288–99.