BMJ Open

Generating high-quality data abstractions from scanned clinical records: text-mining-assisted extraction of endometrial carcinoma pathology features as proof of principle

Anthony Nguyen [iD],[1] John O'Dwyer,[1] Thanh Vu,[1] Penelope M Webb,[2] Sharon E Johnatty,[2] Amanda B Spurdle[2]

[1]The Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Brisbane, Queensland, Australia
[2]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

Correspondence to
Dr Anthony Nguyen;
anthony.nguyen@csiro.au

ABSTRACT

Objective Medical research studies often rely on the manual collection of data from scanned typewritten clinical records, which can be laborious, time consuming and error prone because of the need to review individual clinical records. We aimed to use text mining to assist with the extraction of clinical features from complex text-based scanned pathology records for medical research studies.

Design Text mining performance was measured by extracting and annotating three distinct pathological features from scanned photocopies of endometrial carcinoma clinical pathology reports, and comparing results to manually abstracted terms. Inclusion and exclusion keyword trigger terms to capture leiomyomas, endometriosis and adenomyosis were provided based on expert knowledge. Terms were expanded with character variations based on common optical character recognition (OCR) error patterns as well as negation phrases found in sample reports. The approach was evaluated on an unseen test set of 1293 scanned pathology reports originating from laboratories across Australia.

Setting Scanned typewritten pathology reports for women aged 18–79 years with newly diagnosed endometrial cancer (2005–2007) in Australia.

Results High concordance with final abstracted codes was observed for identifying the presence of three pathology features (94%–98% F-measure). The approach was more consistent and reliable than manual abstractions, identifying 3%–14% additional feature instances.

Conclusion Keyword trigger-based automation with OCR error correction and negation handling proved not only to be rapid and convenient, but also providing consistent and reliable data abstractions from scanned clinical records. In conjunction with manual review, it can assist in the generation of high-quality data abstractions for medical research studies.

Strengths and limitations of this study

► The study presents a rapid and convenient text-mining method to automatically extract pathology features from complex text-based scanned photocopies of typewritten clinical pathology reports drawn from multiple different sources.
► The method can be adapted to address a wide range of textual nuances or artefacts resulting in 'noise' common to scanned PDF images.
► Data quality from text mining methods was validated through the use of statistical significance testing comparing our method to manual abstraction.
► The method can be used in conjunction with manual data abstraction to resolve discrepancies and increase the accuracy of data abstraction.
► The robustness and generalisability of the method are limited to a single medical research study and using a combination of readily available and proven approaches on typewritten reports, as proof of principle.

a critical aspect of cancer research studies. Such data provide confirmatory evidence that patients affected with a specific cancer type meet the diagnostic inclusion criteria for research and clinical studies, and other information important for cancer-related analyses, for example, known prognostic features such as tumour grade and histological subtype, and family history of cancer as relevant for selection for genetic testing.[2] Information about additional features may be collected to enable exploratory research. Overall, manual extraction of pathology information is laborious, time consuming and error prone because of the need to review individual clinical records.[3 4]

Mining electronic health records (EHRs) or electronic medical records (EMRs) using text mining has proven to be an important

INTRODUCTION

Medical research studies often rely on the collection of data from clinical records.[1] Extracting data from pathology reports is

and powerful technique for extracting phenotypic and treatment information about patients.[5 6] Text mining tools that reliably extract features from typewritten pathology reports have been widely developed.[7–17] However, historical paper-based records in the form of photocopied (or scanned) typewritten reports have presented additional challenges for text mining tools, since the optical character recognition (OCR) of individual characters from the scanned images of reports can be error prone. Significant impact of such errors has been reported when text mining tools were applied directly on the raw OCR output of scanned clinical records,[18–20] and degradation in extraction performance has also been reported in the general text mining domain.[21–24]

Techniques for automatically detecting and correcting OCR errors can improve the quality of the OCR text for subsequent interpretation by text mining tools. Common techniques include error pattern matching based on OCR confusions between characters with similar features, for example, the substitution of 'D' for 'O'.[18 25 26] More advanced OCR correction strategies also perform approximate string matching and n-gram analysis.[25 27] Despite the research and development of OCR error correction tools, many clinical and biomedical text mining applications are still processing raw OCR records without error correction.[19 20 28 29]

We developed a simple and convenient text mining tool, coupled with OCR error pattern correction and negation identification, to handle the nuances of scanned records and unstructured pathology reporting. The tool utility was validated on clinical records collected as part of the population-based Australian National Endometrial Cancer Study (ANECS).[30 31] It exemplifies a large-scale cancer research study reliant on manual abstraction of clinical data from paper-based typewritten pathology records, stored as scans of photocopied reports.

As part of a pathology-focused research study assessing the coexistence of leiomyomas, endometriosis and adenomyosis in patients with endometrial cancer (EC) participating in ANECS,[32] the accuracy of manual abstraction of these three pathology features was reviewed by comparing abstractor codes to codes assigned using the text mining tool. It was hypothesised that the text mining tool, using predefined keyword triggers and OCR error corrections and negation handling, would facilitate rapid and accurate data abstractions for clinical research studies.

METHODS
Dataset
ANECS was conducted from 2005 to 2007, and recruited women aged 18–79 years with newly diagnosed EC from across Australia. All ANECS participants provided informed written consent, and approval was obtained from the QIMR Berghofer Medical Research Institute Human Research Ethics Committee, participating hospitals and cancer registries. Details of participant
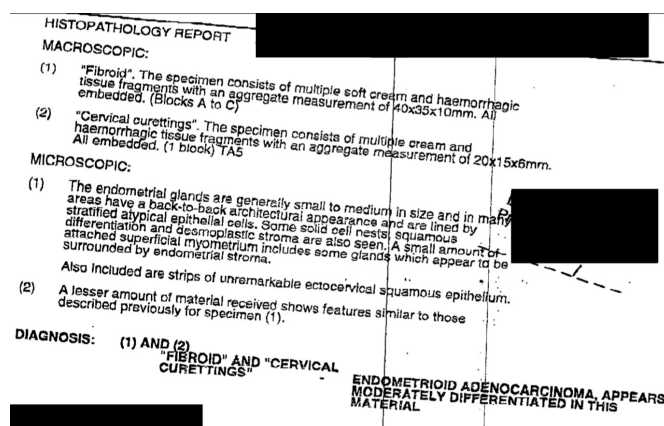

Figure 1 Redacted scanned pathology report.

ascertainment, eligibility criteria, questionnaires and data collection have been previously reported.[30 31]

Photocopies of pathology reports were sent from recruiting sites to the coordinating institution, and scanned for storage as a portable document format (PDF) image. Figure 1 shows an example of a redacted scanned pathology report text used in the study. An abstraction form was developed to facilitate standardised capture of pathology features considered relevant for baseline and exploratory analysis (figure 2).

Pathology reports were reviewed in batches by one of four abstractors: a medical doctor, academic scientist and research nurse (all with extensive experience abstracting information from gynaecological pathology reports), and also by a gynaecological pathologist. Information manually extracted from pathology reports was recorded using hard copies of the abstraction form. The information was then entered into a database using numerical codes. Range and logic checks were performed for key diagnostic and prognostic variables (eg, primary site of cancer, dates of surgery/curette, histological subtype, grade, extent of spread). Formal validation of abstraction, for example, by double abstraction, was not conducted for leiomyomas, adenomyosis and endometriosis. These pathology features were coded as 'Yes', 'No' or 'Not reported'. Specific instructions provided to abstractors allowed abstractors to infer 'No' coding in some instances, namely: If adenomyosis and/or fibroids not specifically mentioned but myometrium is clearly normal then select 'No'. Otherwise select 'Not reported'.

In the parallel ANECS research study,[32] data codes ('Yes', 'No', 'Not reported') for leiomyomas, endometriosis and adenomyosis generated by manual abstraction from diagnostic pathology reports were compared against terms extracted using the text mining tool for 1304 scanned patient reports. Discrepancies were manually cross checked to arrive at a final coding for each feature, based on the presence of terms in the pathology report. Crosschecks were not undertaken to assess if abstractor 'No' and 'Not reported' discrepancies might in fact be abstractor decision to infer that a feature was not present.

Figure 2 Extract of abstraction form for standardised capture of pathology features. FIGO, staging system determined by the International Federation of Gynecology and Obstetrics (Fédération Internationale de Gynécologie et d'Obstétrique); N/A, not applicable; N/R, not reported.

In this study, we detail the methodology and results of the text mining tool evaluation.

Development

Pathology reports were obtained as scanned PDF images. Adobe Acrobat Pro was used to perform OCR on each report to convert the textual information in the PDF images into searchable text.

Inclusion and exclusion search terms (along with spelling variations) were specified, based on expert knowledge, to search for evidence of leiomyomas, endometriosis and adenomyosis (table 1).

The text mining tool (hereinafter called 'system'), a Java-based program, was developed and refined using a random sample set of 11 scanned histopathology reports (hereinafter called the development set). The algorithm was iteratively refined by reviewing discordances between the automated and expected classifications. Refinements included the addition of spelling, synonym and OCR variations found in the scanned pathology reports.

A trial of the system was conducted over another 100 random reports. This was to ensure that the tool output and format were adequate for performing subsequent

Table 1 Inclusion and exclusion search terms selected based on expert knowledge

Evidence type	Leiomyoma	Endometriosis	Adenomyosis
Inclusion search terms	fibroid	endometriosis	adenomyosis
	fibroids		adenomyotic
	leiomyoma		
	leiomyomata		
	leiomyomas		
	smooth muscle		
	neoplasm		
	smooth muscle tumour		
	smooth muscle tumor		
Exclusion search terms	fibrosis	endometritis	
	fibrotic		

crosschecks. No additional modification to the algorithm was deemed necessary after reviewing the output of the 100 reports. The tool was then run over the full dataset.

The system reads in OCRed PDF files containing the pathology reports, and a configuration file containing the list of user-specified search terms. Coded abstracted data in tabular comma-separated values (CSV) format were then output, detailing the file name and the coded output for each pathology feature.

The configuration file was in two sections. The first section was a list of search terms with their corresponding coding (eg, 'leiomyoma' is coded as a 'Yes' value; 'no_leiomyoma' is coded as a 'No' value; 'not_leiomyoma' is an exclusion term that would be ignored). The second part dealt with the 'noisy' nature of scanned PDF images. Search terms were expanded with their character variations based on common OCR error patterns (eg, 'i' can be 'l' or '!'; and 'm' can also be 'rn') identified in the development set. The configuration file included the list of possible OCR error patterns to consider (eg, 'i–>l|!' and 'm–>rn'; where 'i' and 'm' could be optionally substituted with 'l' or '!' and 'rn', respectively).

The system used the configuration file parameters to create regular expression (regex) patterns—a widely adopted technique in text processing—for each search term.[33 34] Regular expressions were used to define a special text string to describe the search patterns for extracting each of the search terms. OCR error patterns in regex were represented within parentheses ('()') with a vertical bar ('l') to separate each character option. Table 2 shows examples of search terms, their corresponding regex search pattern and the textual context from the PDF report that contained the search term.

The matching of search terms was case insensitive and based on the longest possible text string match starting from any position. For each PDF report, pathology features were assigned one of three values ('Yes', 'No', 'Not reported'). If search terms were identified in the scanned report, then the corresponding feature was assigned a 'Yes' value. A couple of negated assertion phrases containing the search terms were also added to the set of search patterns (eg, 'adenomyosis: absent' with 'absent' as a search term value and 'no adenomyosis' with 'no' appearing immediately before the search term). If these phrases were identified in the PDF report, then the corresponding feature was assigned a 'No' value. If conflicting values were found by the tool for a given PDF report (ie, 'Yes' and 'No' values for a given feature), then the tool would output a 'Conflict' value. The decision to introduce the option of 'Conflict' in the system allowed such cases to be revisited and manually resolved. However, if no search terms were found or were only found from the 'exclusion' list, then the patient was assigned a 'Not reported' value. The system also output an additional CSV file to detail the sentence context surrounding each search term found in the PDF report, to assist with quality assurance checks (eg, additional manual crosschecking of discrepancies).

Evaluation

The system output was crosschecked against the original abstracted coding and differences were resolved to generate the final curated dataset. If a term was identified by the system, but the pathology abstraction code was 'No'/'Not reported', then the extracted sentence context provided by the system was reviewed. If a feature was identified by pathology abstraction but not identified by the system, then the entire pathology report was rereviewed for evidence of the given feature. At this time, crosschecks were done for the remaining two features, so providing additional confirmatory review of concordant results in parallel. Overall, 589 records were reviewed for a combination of discordant and concordant results; 281 pathology reports were manually reviewed in full, while the system-generated context around a specific term was

Table 2 Example search terms, regular expression search patterns and textual context in the portable document format report containing the search term (shown in *italics*)

Search term	Regular expression pattern	Textual context
leiomyoma	(l|i)e(i|l|!)(o|a|c)(m|rn)y(o|a|c)(m|rn)(a|o)	"myometr!um contains a benign *leiomyorna*"
endometriosis	end(o|a|c)(m|rn)etr(i|l|!)(o|a|c)s(i|l|!)s	"right fallopian tube shows a focus of *endometriosis*"
adenomyosis: absent	(a|o)den(o|a|c)(m|rn)y(o|a|c)s(i|l|!)s: (a|o)bsent	"evidence of *adenomyosls: absent*"

checked to confirm or revise coding for the remaining 308 records. The final set of abstracted codes obtained from the manual crosschecking of discrepancies was used as the gold standard for evaluations.

A contingency table for each of the pathology features was used to tabulate frequency counts of 'Yes', 'No', and 'Not reported' values assigned by the system/abstractor and final abstracted codes. This was used to assess concordance between the system/abstractor and final abstracted codes, as well as the impact from abstractor inferences in the coding of 'No' cases for leiomyomas and adenomyosis (see the Dataset section).

To evaluate the effectiveness of the tool, positive predictive value (PPV), sensitivity and F-measure (a single, overall evaluation measure representing the harmonic mean of PPV and sensitivity) were reported on the non-developmental set of reports (hereinafter called the evaluation set).[35] For evaluation purposes, 'Conflict' values output by the system were considered a 'No' classification as specific evidence for a negated feature was found within the pathology report. The contribution of OCR error correction and negation handling on the performance of the system was also assessed.

The statistical significance between the difference in performances between the system and abstractor, as well as across the different system configuration settings, was established using the approximate randomisation test,[36 37] with n=9999 and significance level alpha of 0.05 and 0.01—representing significant and very significant differences, respectively. The approximate randomisation test is a standard non-parametric statistical significance test for text mining tasks.[36 37]

Patient and public involvement

No patients and/or public were involved during identifying the research question or during the design and conduct of the study.

RESULTS

The final coded abstraction statistics for the three pathology features after manual crosschecking of discrepancies between the system and abstractor is shown in table 3.

Contingency tables detailing the matches for the system (with OCR correction and negation handling) and abstractor against the final set of abstracted codes are shown in table 4. Results along the main diagonal (bold font) show feature value concordance, while the off-diagonal results show the feature value discrepancies.

There were seven cases of 'Conflict' output by the system indicating both the presence and negated assertion of a pathology feature being found in the same report. These 'Conflict' values allowed for corresponding cases to be revisited and manually resolved. The decision on the final coding for these cases depended on the context of its mention in the report, and thus could result in a coding of a 'Yes' or 'No' value (see table 4). As before-mentioned in the Evaluation section, 'Conflict' values were considered a 'No' for the purposes of evaluations. The larger discrepancies in abstractor coding of 'No' and 'Not reported' values for adenomyosis and leiomyomas also highlight the possible extent of abstractor inference in the coding of 'No' values.

As the ANECS pathology-focused research study on leiomyomas, endometriosis and adenomyosis analysed the coexistence (and thus the 'presence') of these conditions,[32] the gold standard was subsequently formulated as binary feature values of 'Yes' and 'Other' (ie, 'No' and 'Not reported' collapsed) for evaluations.

Table 5 presents the performance of the system and abstractors in coding a 'Yes' or 'Other' value for each pathology feature. Overall, based on F-measure, the system achieved higher performances than abstractors for all three pathology features. Across all the evaluation metrics, system performances were either consistently competitive (no statistically significant difference) or statistically significantly better than abstractor.

Table 6 presents the contribution of OCR error correction and negation handling on the performance of the system. The baseline system results, using exact match of search terms, showed very strong performances. Negation handling provided significant improvements over the baseline system approach for leiomyomas and adenomyosis. Incremental improvements on top of negation handling were observed when OCR correction was applied, except for endometriosis where no additional terms requiring correction were identified.

DISCUSSION

The system was observed to have very high concordance against final coding (at least 94.5% F-measure), demonstrating consistent and reliable extractions across all pathology features. This resulted in identifying an additional 3%–14% of the number of 'Yes' feature values when compared with manual abstractions (9.6% increase for leiomyomas, 14.4% for endometriosis and 3.6% for adenomyosis). The additional features identified by the system allowed for a more accurate dataset to be curated.

The use of readily available and proven OCR and text mining approaches, in combination, proved to be highly effective. The combination of expert knowledge (ie, specification of search terms) with the small number of example cases to extrapolate textual patterns across

Table 3 Final coded abstraction statistics for leiomyomas, endometriosis and adenomyosis

Pathology feature	Final abstracted coding (development/ evaluation set)		
	Yes	No	Not reported
Leiomyomas	693 (9/684)	25 (0/25)	586 (2/584)
Endometriosis	106 (1/105)	14 (0/14)	1184 (10/1174)
Adenomyosis	538 (5/533)	36 (3/33)	730 (3/727)

Table 4 Contingency table for system/abstractor and the final abstracted codes on the evaluation set for (a) leiomyomas, (b) endometriosis and (c) adenomyosis

	System				Abstractor		
	Yes	**No**	**Not reported**	**Conflict**	**Yes**	**No**	**Not reported**
Final abstracted codes (n)							
(a) Leiomyomas							
Yes (684)	**673**	1	9	1	**614**	22	48
No (25)	8	**16**	0	1	1	**21**	3
Not reported (584)	5	0	**579**	0	12	196	**376**
Total (1293)	686	17	588	2	627	239	427
(b) Endometriosis							
Yes (105)	**103**	0	0	2	**90**	1	14
No (14)	10	**3**	0	1	1	**3**	10
Not reported (1174)	0	0	**1174**	0	2	14	**1158**
Total (1293)	113	3	1174	3	93	18	1182
(c) Adenomyosis							
Yes (533)	**515**	0	18	0	**497**	15	21
No (33)	7	**24**	0	2	3	**29**	1
Not reported (727)	2	0	**725**	0	5	252	**470**
Total (1293)	524	24	743	2	505	296	492

Results along the main diagonal (bold font) show feature value concordance, while the off-diagonal results show the feature value discrepancies.
Discrepancies in abstractor coding of 'No' and 'Not reported' values for leiomyomas and adenomyosis highlights the possible extent of abstractor inference in the coding of 'No' values.

pathology features and feature values was also key for developing a high performing system.

The incorporation of negation handling proved to have a significant impact on the results. Negation handling reduced the number of false-positive search terms that would have otherwise been found. The system miscoding of 'No' cases was observed to be caused by negative assertion phrases that were not specified in the system.

Although the system could incorporate more robust negation detectors,[38 39] performing error analysis to specify additional negation phrases could result in immediate gains with minimal effort.

The OCR error correction technique based on regular expressions proved to be effective at detecting search terms in the presence of OCR errors. Although improvements on top of negation handling due to OCR

Table 5 System effectiveness results for leiomyomas, endometriosis and adenomyosis classification on the evaluation set

	Yes			Other		
	PPV	**Sensitivity**	**F-measure**	**PPV**	**Sensitivity**	**F-measure**
Leiomyomas						
Abstractor	97.93%	89.77%	93.67%	89.49%	97.87%	93.49%
System	98.11%	98.39%†	98.25%†	98.19%†	97.87%	98.03%†
Endometriosis						
Abstractor	96.77%	85.71%	90.91%	98.75%	99.75%	99.25%
System	91.15%	98.10%†	94.50%	99.83%†	99.16%	99.49%
Adenomyosis						
Abstractor	98.42%	93.25%	95.76%	95.43%	98.95%	97.16%
System	98.28%	96.62%*	97.45%*	97.66%*	98.82%	98.23%

*Performance difference between system and abstractor is significant at alpha = 0.05.
†Performance difference between system and abstractor is very significant at alpha = 0.01.
PPV, positive predictive value.

Table 6 Contribution of optical character recognition (OCR) error correction and negated assertions on the performance of the system

	Yes			Other		
	PPV	Sensitivity	F-measure	PPV	Sensitivity	F-measure
Leiomyomas						
Baseline	95.86%	98.25%	97.04%	97.97%	95.24%	96.59%
+Negated assertions	98.10%*	97.95%	98.03%*	97.71%	97.87%*	97.79%*
+OCR correction	98.11%*	98.39%	98.25%*	98.19%	97.87%*	98.03%*
Endometriosis						
Baseline	88.98%	100.00%	94.17%	100.00%	98.91%	99.45%
+Negated assertions	91.15%	98.10%	94.50%	99.83%	99.16%	99.49%
+OCR correction	91.15%	98.10%	94.50%	99.83%	99.16%	99.49%
Adenomyosis						
Baseline	93.59%	95.87%	94.72%	97.06%	95.40%	96.22%
+Negated assertions	98.27%*	95.87%	97.06%*	97.15%*	98.82%*	97.98%*
+OCR correction	98.28%*	96.62%	97.45%*	97.66%*	98.82%*	98.23%*

Baseline configuration refers to the exact match of search terms.
*Performance difference against baseline is very significant at alpha = 0.01.
PPV, positive predictive value.

error correction were not significant, the configuration allowed for the detection of additional features that may have been missed by both the exact match approach and abstractors. The value of OCR error correction is dependent on the quality of the OCR software employed and the type of artefacts present in the scanned versions of the pathology reports.[19]

The system is highly configurable and allows for additional search patterns to be specified. The rereview of discordant cases could be analysed to identify additional search patterns. Additional search patterns may include new OCR error patterns and writing variations such as medical shorthand notations. On rereview of system 'Yes' cases where the gold standard was 'Other', it was observed that questions marks preceding search terms, indicating a feature to be investigated, generated many false positives (eg, '?endometriosis'). Such a search term pattern can be easily specified as an exclusion search term to generate more accurate results.

Abstractor coding discrepancies were mainly related to the differences in coding of 'Not reported' versus 'No', which for leiomyomas and adenomyosis (but not for endometriosis) were likely to have been at least partly due to abstractor inference that a feature was not present, based on abstraction instructions provided (see the Dataset section). More sophisticated text mining techniques have the potential to perform inferences, and would be a promising avenue of future work.[15 17] Other abstractor coding errors were due to the manual and subjective nature of the abstraction task where the presence (or mention) of pathology features in reports was overlooked by the abstractor.

In general, system and abstractor errors were found to be attributed to poor quality of the scanned reports.

Search terms were sometimes not picked up by either the abstractor or system because of poor scan quality or background 'noise' such as random markings through the text.

Although errors were inevitable by either the system and/or abstractor, the automatic extraction of information from scanned pathology reports was invaluable in identifying and resolving discrepancies between the system and abstractors. The adjudication process greatly enhanced the accuracy of the ANECS pathology dataset for the analysis of the coexistence of leiomyomas, endometriosis and adenomyosis features in EC.[32] Though the system was applied to a single medical research study as proof of principle, its robustness and generalisability in other medical research studies will need to be determined.

Despite the availability of EMRs that store electronic text, a substantial proportion of current and historical records are still available in scanned PDF image formats. These scanned medical records can be either handwritten or typewritten. The work and literature reported in this study were concerned with typewritten documents. Further studies would be necessary to evaluate the role of the proposed system on handwritten documents, as the OCR of handwritten documents can be more challenging.[40]

The proposed system with OCR correction capability and negation handling has broad applicability and could be applied in clinical settings and specialised clinical studies for the extraction of other clinical conditions (or phenotypes) and biological entities to create searchable databases of medical records and/or biomedical literature from scanned document archives.[26 28 41] Other applications of text mining on scanned medical records can extend to health business intelligence and health

service improvements activities such as patient recruitment in clinical studies,[26] cancer registry coding[19] and the processing of patient referrals.[29]

CONCLUSION

A text mining tool based on search term trigger-based automation with OCR error correction and negation handling was highly accurate in extracting information from scanned textual medical records. It greatly enhanced the curation of a manually abstracted pathology research dataset. The value of this approach was demonstrated to reliably extract and code equivalent terms from scanned medical records for the text-mining-assisted generation of clinical datasets.

ORCID iD
Anthony Nguyen http://orcid.org/0000-0002-6215-6954

REFERENCES

1 Nordo AH, Levaux HP, Becnel LB, *et al*. Use of EHRs data for clinical research: historical progress and current applications. *Learn Health Syst* 2019;3:e10076.
2 Edwards E, Lucassen A. The impact of cancer pathology confirmation on clinical management of a family history of cancer. *Fam Cancer* 2011;10:373–80.
3 Burger G, Abu-Hanna A, de Keizer N, *et al*. Natural language processing in pathology: a scoping review. *J Clin Pathol* 2016;69:949–55.
4 Meystre SM, Lovis C, Bürkle T, *et al*. Clinical data reuse or secondary use: current status and potential future progress. *Yearb Med Inform* 2017;26:38–52.
5 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13:395–405.
6 Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34–49.
7 Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: a review. *JAMA Oncol* 2016;2:797–804.
8 Coden A, Savova G, Sominsky I, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *J Biomed Inform* 2009;42:937–49.
9 Carrell DS, Halgrim S, Tran D-T, *et al*. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol* 2014;179:749–58.
10 Savova GK, Danciu I, Alamudun F, *et al*. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res* 2019;79:5463–70.
11 Martinez D, Li Y. *Information extraction from pathology reports in a hospital setting*. Glasgow, Scotland, UK: Proc ACM Int Conf Inf Knowl Manag, 2011: 1877–82.
12 Ou Y, Patrick J. Automatic structured reporting from narrative cancer pathology reports. *J Health Inform* 2014;8:e20.
13 Currie A-M, Fricke T, Gawne A, *et al*. Automated extraction of Free-Text from pathology reports. *AMIA Annu Symp Proc* 2006;899.
14 Buckley JM, Coopey SB, Sharko J, *et al*. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J Pathol Inform* 2012;3:23.
15 Nguyen AN, Lawley MJ, Hansen DP, *et al*. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc* 2010;17:440–5.
16 Nguyen AN, Moore J, O'Dwyer J, *et al*. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. *AMIA Annu Symp Proc* 2015;2015:953–62.
17 Khor RC, Nguyen A, O'Dwyer J, *et al*. Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. *Int J Med Inform* 2019;121:53–7.
18 Li X, Zhang D, Liu B. Automated extraction of radiation dose information from CT dose report images. *AJR Am J Roentgenol* 2011;196:W781–3.
19 Zuccon G, Nguyen AN, Bergheim A, *et al*. The impact of OCR accuracy on automated cancer classification of pathology reports. *Stud Health Technol Inform* 2012;178:250–6.
20 Zuccon G, Kotzur D, Nguyen A, *et al*. De-Identification of health records using Anonym: effectiveness and robustness across datasets. *Artif Intell Med* 2014;61:145–51.
21 Taghva K, Nartker T, Borsack J. Information access in the presence of OCR errors. ACM workshop on Hardcopy document processing. *ACM* 2004:1–8.
22 Miller D, Boisen S, Schwartz R, *et al*. *Named entity extraction from noisy input: speech and OCR*. ANLC '00: Proceedings of the sixth conference on Applied natural language processing, 2000: 316–24.
23 Grover C, Givon S, Tobin R, *et al*. *Named entity recognition for digitised historical texts*. LREC: European Language Resources Association (ELRA), 2008.
24 Rodriquez KJ, Bryant M, Blanke T, *et al*. *Comparison of named entity recognition tools for RAW OCR text*. KONVENS, 2012: 410–4.
25 Kukich K. Technique for automatically correcting words in text. *ACM Comput Surv* 1992;24:377–439.
26 Jackson R, Kartoglu I, Stringer C, *et al*. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. *BMC Med Inform Decis Mak* 2018;18:47.
27 Taghva K, Stofsky E. OCRSpell: an interactive spelling correction system for OCR errors in text. *IJDAR* 2001;3:125–37.
28 Ehsani S, Kiehl T-R, Bernstein A, *et al*. Creation of a retrospective searchable neuropathologic database from print archives at Toronto's university health network. *Lab Invest* 2008;88:89–93.
29 Todd J, Richards B, Vanstone BJ, *et al*. Text mining and automation for processing of patient referrals. *Appl Clin Inform* 2018;9:232–7.

30 Rowlands IJ, Nagle CM, Spurdle AB, *et al*. Gynecological conditions and the risk of endometrial cancer. *Gynecol Oncol* 2011;123:537–41.

31 Johnatty SE, Tan YY, Buchanan DD, *et al*. Family history of cancer predicts endometrial cancer risk independently of Lynch syndrome: implications for genetic counselling. *Gynecol Oncol* 2017;147:381–7.

32 Johnatty SE, Stewart CJR, Smith D, *et al*. Co-existence of leiomyomas, adenomyosis and endometriosis in women with endometrial cancer. *Sci Rep* 2020;10:3621.

33 Aho AV. *Compilers: principles, techniques and tools, 2/e*. Pearson Education India, 2003.

34 Stubblebine T. *Regular expression pocket reference: regular expressions for Perl, Ruby, PHP, Python, C, Java and NET*. O'Reilly Media, Inc, 2007.

35 Hripcsak G, Rothschild AS, Agreement RAS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296–8.

36 Chinchor N. *The statistical significance of the MUC-4 results*. Proceedings of the 4th conference on message understanding. Association for Computational Linguistics, 1992: 30–50.

37 Dror R, Baumer G, Shlomov S, *et al*. *The hitchhiker's guide to testing statistical significance in natural language processing*. . Proc Conf Assoc Comput Linguist Meet, 2018: 1. 1383–92.

38 Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34:301–10.

39 Nguyen AN, Lawley MJ, Hansen DP. *A simple pipeline application for identifying and negating SNOMED clinical terminology in free text*. HIC, 2009: 188–96.

40 Tafti AP, Baghaie A, Assefi M, *et al*. *OCR as a service: an experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym*. International Symposium on Visual Computing, 2016: 735–46.

41 Westergaard D, Stærfeldt H-H, Tønsberg C, *et al*. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding Abstracts. *PLoS Comput Biol* 2018;14:e1005962.

9