INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 83 (2014) 605–623

journal homepage: www.ijmijournal.com

Review

Text mining of cancer-related information: Review of current status and future directions

Irena Spasić[a,*], *Jacqueline Livsey*[b], *John A. Keane*[c,d,e], *Goran Nenadić*[c,d,e]

[a] *School of Computer Science & Informatics, Cardiff University, Cardiff CF24 3AA, UK*
[b] *Clinical Outcomes Unit, The Christie NHS Foundation Trust, Manchester M20 4BX, UK*
[c] *School of Computer Science, The University of Manchester, Manchester M13 9PL, UK*
[d] *Health e-Research Centre, Manchester M13 9PL, UK*
[e] *Manchester Institute of Biotecnology, Manchester M1 7DN, UK*

ARTICLE INFO

Article history:
Received in revised form
12 June 2014
Accepted 14 June 2014

Keywords:
Cancer
Natural language processing
Data mining
Electronic medical records

ABSTRACT

Purpose: This paper reviews the research literature on text mining (TM) with the aim to find out (1) which cancer domains have been the subject of TM efforts, (2) which knowledge resources can support TM of cancer-related information and (3) to what extent systems that rely on knowledge and computational methods can convert text data into useful clinical information. These questions were used to determine the current state of the art in this particular strand of TM and suggest future directions in TM development to support cancer research.

Methods: A review of the research on TM of cancer-related information was carried out. A literature search was conducted on the Medline database as well as IEEE Xplore and ACM digital libraries to address the interdisciplinary nature of such research. The search results were supplemented with the literature identified through Google Scholar.

Results: A range of studies have proven the feasibility of TM for extracting structured information from clinical narratives such as those found in pathology or radiology reports. In this article, we provide a critical overview of the current state of the art for TM related to cancer. The review highlighted a strong bias towards symbolic methods, e.g. named entity recognition (NER) based on dictionary lookup and information extraction (IE) relying on pattern matching. The F-measure of NER ranges between 80% and 90%, while that of IE for simple tasks is in the high 90s. To further improve the performance, TM approaches need to deal effectively with idiosyncrasies of the clinical sublanguage such as non-standard abbreviations as well as a high degree of spelling and grammatical errors. This requires a shift from rule-based methods to machine learning following the success of similar trends in biological applications of TM. Machine learning approaches require large training datasets, but clinical narratives are not readily available for TM research due to privacy and confidentiality concerns. This issue remains the main bottleneck for progress in this area. In addition, there is a need for a comprehensive cancer ontology that would enable semantic representation of textual information found in narrative reports.

* *Corresponding author*. Tel.: +44 029 2087 0320.
E-mail address: i.spasic@cs.cardiff.ac.uk (I. Spasić).
http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009

Contents

1. Introduction

Around 325,000 people were diagnosed with cancer in 2010 in the UK (approximately 890 people per day) [1]. More than 1 in 3 people in the UK will develop some form of cancer during their lifetime. Worldwide, around 12.7 million new cases of cancer were estimated in 2008. Currently, half of people diagnosed with cancer will survive for at least five years. With constant advancements in cancer research and healthcare provision, cancer survival rates in the UK have doubled in the last 40 years. Still, cancer remains the most common cause of death (29%) followed by circulatory diseases (28%) such as heart disease and strokes [2]. Further advances depend crucially on consistent and comparable data collection.

A wealth of relevant information exists in various types of medical records, e.g. approximately 96% of cancer diagnoses originate in the surgical pathology laboratory [3] and as such they remain an important source for information to guide the treatment of patients with cancer. Synoptic reports emerged with the goal of structured data capture as means of improving the quality of data and collection methods. The College of American Pathologists (CAP) as part of their contribution to the International Collaboration on Cancer Reporting developed cancer checklists that prescribe collection of all critical elements that should be reported for cancer specimens [4]. Still, in many cases (e.g. colonoscopy) free-text reporting via dictation is still the norm [5]. While synoptic reporting systems can effectively address structured data collection in the future, it does not solve the problem of managing the legacy data in free text format.

Text mining (TM) has emerged as a potential solution for bridging the gap between free-text and structured representation of cancer information. It uses techniques from natural language processing (NLP), knowledge management, data mining and machine learning (ML) to efficiently process large document collections in order to support information retrieval (which gathers and filters relevant documents), document classification (which maps documents to appropriate categories based on their content), information extraction (which selects specific facts about pre-specified types of entities and relationships of interest), terminology extraction (which collects domain-relevant terms from a corpus of domain-specific documents), named entity recognition (which identifies entities from predefined categories), etc.

A recent article provides a review of TM applications in cancer research [6]. It describes biomedical TM in general terms and how it can be used in cancer systems biology. Systems biology studies complex interactions in biological systems such as gene regulatory networks, which are often explored to provide insight into the cell proliferation observed in cancer. Therefore, most of the TM approaches discussed in the mentioned review article focus on PubMed abstracts as the richest source of text data on gene interactions and NLP techniques to extract gene-related information. In this article, we depart from biological applications of TM to focus primarily on its clinical applications in the cancer domain. To address the interdisciplinary nature of such research, we searched the scientific literature encompassing the following areas: biomedicine, life sciences, engineering, technology, computing and information technology. We relied on three literature databases: Medline, IEEE Xplore and ACM digital libraries. The search results were supplemented with the literature identified through Google Scholar. The search terms included *cancer* and related terms (e.g. *carcinoma*, *neoplasm*, *malignancy*, etc.) in combination with terminology related to *text mining* (e.g. *natural language processing*, *information retrieval*, *information extraction*, etc.). To be included in the review, the citations were required to discuss text mining of cancer-related information demonstrated on or directly applicable to clinical narratives,

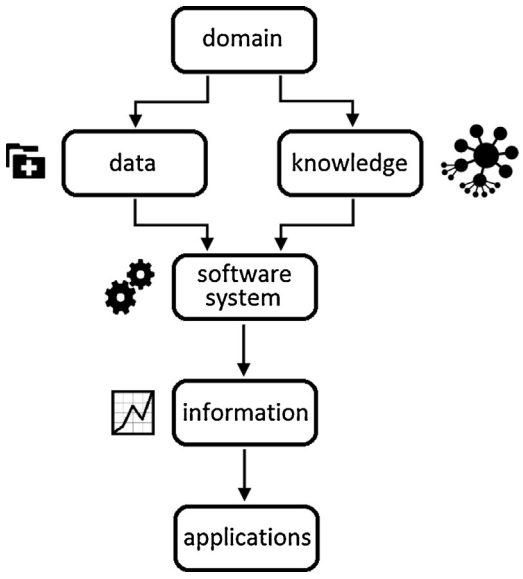

Fig. 1 – Text mining framework for clinical applications.

which was determined by manually screening the full text of retrieved publications.

Fig. 1 presents the general framework in which TM is used in the clinical setting. Different types of cancer (e.g. breast cancer) will define a specific domain in which TM is used. It will determine the choice of available text data (e.g. mammography reports). Data interpretation, either by human experts or computers, naturally requires relevant knowledge in a given domain. Machine readable representations of relevant knowledge can be found in generic knowledge representation systems (e.g. UMLS [7]), specialised databases (e.g. BCGD [8]), reporting guidelines (e.g. BI-RADS [9]), etc. Specialised TM software systems are then used to extract information (e.g. cancer stage) as a joint function of data and knowledge. Extracted information can further support various applications including clinical research (e.g. epidemiology) and decision making (e.g. the choice of appropriate treatment). Before automatically extracted information can be used in clinical setting, its reliability needs to be validated. Analogously to establishing the validity of diagnostic tests, TM systems are evaluated in terms of a range of measures such as positive and negative predictive value to determine to what extent the extracted information corresponds to the reality represented by the ground truth.

The TM framework given in Fig. 1 has been used to provide a structure of this review. In Section 2, we discuss which cancer domains have been the subject of TM efforts and which ones could be suitable targets of TM research in the near future. A specific cancer domain will constrain the types of available text data, which are discussed in Section 3. The available knowledge resources that can support TM of cancer-related information are summarised in Section 4. Section 5 describes TM systems that rely on knowledge and computational methods to convert text data into useful clinical information. To ascertain the feasibility of using TM to extract different types of cancer-related information, this section also features the means of formally evaluating TM systems. The presented systems are discussed in the context of their applications. Finally, we conclude the paper with a discussion of the current state of the art in this particular strand of TM and suggest future directions in TM development to support cancer research.

Table 1 – Examples of text mining studies by cancer domain.

Neoplasm by site	CUI	Study
Breast neoplasm	C1458155	[14–21]
Cervical neoplasm	C0007873	[22]
Colon neoplasm	C0009375	[23,24]
Colorectal neoplasm	C0009404	[5,25–30]
Lung neoplasm	C0024121	[26,31,32]
Ovarian neoplasm	C0919267	[33]
Pancreatic neoplasm	C0030297	[34]
Prostate neoplasm	C0033578	[21,26,35–37]
Skin neoplasm	C0037286	[36]

2. Cancer domains

Cancer is an umbrella term for diseases characterised by excessive cellular division and proliferation. Cancer is no longer viewed as a single disease or even a single collection of diseases. Cancer may start developing in any of over 60 body organs and is usually named after the affected organ (e.g. breast cancer). Each organ consists of different cell types that may be affected by cancer, so several cancer types may affect each organ (e.g. ductal carcinoma and lobular carcinoma are different types of breast cancer). There are more than 200 types of cancer having different causes, symptoms and treatments. Therefore, cancer as a term covers a diverse set of diseases, which are differentiated by a growing set of biomarkers. Naturally, different types of information will be reported for different types of cancer, so it may be expected that TM approaches to processing text data in different cancer domains will vary as well. Nonetheless, it would be useful to know what knowledge and software resources can be re-used in a specific domain. We, therefore, provide a brief summary of TM activities across different cancer types. Detailed information about the actual approaches taken is provided in Section 5.

The literature review revealed that most of the articles on TM for cancer have been annotated with a MeSH term that identifies neoplasms by anatomical site, where neoplasm refers to autonomous tissue growth in which the malignancy status has not been established and for which the transformed cell type has not been specifically identified. Both NCI Thesaurus [10] and MeSH [11] organise neoplasms by anatomical site. Unified Medical Language System (UMLS) [7] currently integrates 168 vocabularies, including both NCI Thesaurus and MeSH. We therefore provide a unique concept identifier (CUI) in UMLS, from which cross-references can be obtained to both of these sources. Information from these knowledge sources can be used to obtain lexical information (e.g. synonyms) as well as links to other related concepts. For practical purposes, both resources are accessible via the UMLS Terminology Services [12] and BioPortal [13].

Table 1 provides the mappings of TM studies to particular cancer domains. Breast, lung, bowel and prostate cancers

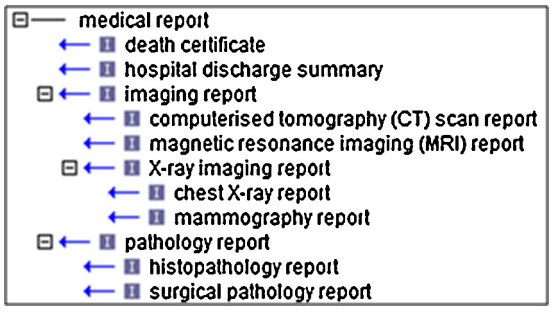

Fig. 2 – A hierarchy of medical reports.

together account for over half of all new cancers each year [1]. Not surprisingly, most TM efforts have concentrated on these domains. While most common cancers have been covered, others such as bladder neoplasm, endometrial neoplasm, renal neoplasm and thyroid gland neoplasm still remain to be explored by TM. In addition, we have not identified any TM studies related specifically to haematologic neoplasms (e.g. leukaemia or lymphoma), which are located in the blood and blood-forming tissue (i.e. the bone marrow and lymphatic tissue).

3. Data sources

Cancer-related information is described in various types of text documents including scientific literature, medical records, web documents and those found in specialised databases. In this section, we focus on different types of text data used in TM studies to support cancer research. Most research in biomedical TM, including mining of cancer-related information, has been conducted on scientific literature. The availability of article abstracts via the PubMed database has made them the most popular source (e.g. [38]). While PubMed abstracts ensure the breadth of information that can be mined, they do not provide for depth, i.e. more detailed information can only be found in the body of the article and some studies have explored this information space with TM (e.g. [34]). Unfortunately, access restrictions limit the availability of full-text articles. Additionally, unlike the abstracts available in a single resource such as PubMed, which have a uniform format, full text articles will be in different formats, which poses difficulties in their processing. PubMed Central is an attempt to make full-text articles freely available in a uniform format, but it currently has a limited coverage compared to PubMed. In addition to literature, much information relevant to cancer available in free text records in specialised databases has been used in TM studies, e.g. AERS (Adverse Event Reporting System) [39], ClinicalTrials.gov [40] and HPV Sequence Database [41].

In a clinical setting, cancer-related information can be found in various types of medical reports. For the purposes of TM, we focus on electronic medical records (EMRs) – electronic documents that may describe demographic information, medical history, medication and known allergies, laboratory test results, radiology images, etc. Fig. 2 shows a hierarchy of medical reports that have been used as data sources in TM for cancer. In particular, two types of reports are relevant for recording cancer-related information: pathology and imaging reports. A pathology report describes the results of examining cells and tissues under a microscope following a biopsy or surgery [42]. It typically contains information about the patient, a description of how cells look under the microscope and a diagnosis. This information is then used by clinicians to support decision making on appropriate treatment.

Table 2 – Examples of document types used in text mining studies across different cancer types.

Cancer type	Document type	Study
Breast neoplasm	Mammography reports	[14,17–19]
Breast neoplasm	Pathology reports	[16,20,21]
Breast neoplasm	PubMed abstracts	[15]
Cervical neoplasm	PubMed abstracts	[22]
Colon neoplasm	Pathology reports	[23,24]
Colorectal neoplasm	EMR notes	[25,26,28,29]
Colorectal neoplasm	Pathology reports	[27]
Colorectal neoplasm	Histopathology reports	[30]
Colorectal neoplasm	Colonoscopy reports	[5]
Lung neoplasm	Radiographic reports	[31]
Lung neoplasm	EMR	[26]
Lung neoplasm	Pathology reports	[32]
Ovarian neoplasm	GPRD records	[33]
Pancreatic neoplasm	PubMed abstracts, EMRs	[34]
Prostate neoplasm	Clinical records: all available paper, electronic, radiologic, radiation therapy and pathology records	[37]
Prostate neoplasm	Pathology reports	[21,36]
Prostate neoplasm	EMR	[26]
Skin neoplasm	Pathology reports	[36]

Imaging reports (or radiology reports) serve the same purpose of conveying a specialist interpretation of images and relate it to the patient's signs and symptoms in order to suggest diagnosis [43]. Depending on the type of imaging technique used, we can further differentiate between different subclasses of imaging reports each having its own reporting standards or guidelines, e.g. X-ray imaging reports (including chest radiography and mammography reports), computed tomography (CT) scan reports and magnetic resonance imaging (MRI) reports. Table 2 shows which types of reports have been used in TM studies focusing on different cancer types.

Hospital discharge summaries may contain cancer-related information among other health-related topics. An anonymised data set was initially released for the i2b2 NLP challenge [44] and it can be used for research purposes (e.g. [37]). Finally, medical certificate of death is a document issued by a medical practitioner explaining the cause of death. Cancer monitoring and prevention rely on timely notification of cancer-related events including deaths. Cancer Registries systematically collect such information, but they may lag behind due to manual classification of cancer from the free-text documents such as death certificates, which is complex and time-consuming activity [45].

While cancer research stands to benefit from automated processing of EMRs, there are ethical and legal issues associated with their use as they contain private and confidential information [46]. Patient data cannot be used for TM purposes without consent [47]. In the EU, Data Protection Directive

(officially Directive 95/46/EC) [48] specifies rules on using information about individuals. In the USA, the Health Insurance Portability and Accountability Act (HIPAA) [49] provides the standard for using patient data in electronic format. Their violations may incur legal responsibilities and penalties. In principle, no patient data should be individually identifiable. This applies to anonymous data (collected without patient-identifiable information), anonymized data (patient-identifiable information is removed) and de-identified data (patient-identifiable information is encoded or encrypted) [50]. Unfortunately, even when all reasonable measures to protect privacy and confidentiality are taken (e.g. de-identification and data use agreements), in most cases only researchers with local affiliation are allowed access [51]. This remains the main bottleneck, more widely, for progress in healthcare applications of NLP. Most systems described in this review were developed and evaluated on local data, which does not allow for their direct comparison or for estimating how generalisable their methods are.

4. Knowledge sources

To identify different cancer patient cohorts from their medical records or obtain baseline classification for automated methods, many studies relied on the relevant coding systems used. In the USA, billing data typically consists of codes derived from the International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) [52]. ICD is a taxonomy of diseases, signs, symptoms and procedure codes maintained by the World Health Organisation (WHO). For example, Coden et al. used ICD-9-CM codes (153.x, 154.x) to select pathology reports of patients diagnosed with colon cancer [23]. Similarly, D'Avolio et al. used ICD-9 codes to select pathology reports related to colorectal cancer (153.x, 154.x), prostate cancer (185.x) and lung cancer (162.x) [26]. ICD version 9-CM is in use in the USA (as of 2012), whereas the majority of other countries use ICD version 10. For example, Butt et al. relied on death classifications based on ICD-9 and ICD-10 codes that accompany the death certificate reports in Australia [45]. More detailed classification of oncology information is available via the International Classification of Diseases – Oncology (ICD-O) version 3, a domain-specific extension of ICD, which is considered as the lingua franca of pathologists and is in widespread use within tumour registries [23]. In PubMed, relevant literature can be selected using MeSH classifications [11]. Most articles relevant for cancer will be annotated with a term from the Neoplasms (C04) section.

TM of cancer data requires understanding of the underlying terminology, which is described in the NCI Thesaurus [10], a reference terminology of the cancer domain. The NCI Thesaurus provides definitions and synonyms of nearly 10,000 cancers and related diseases, 8000 single agents and combination therapies and a range of other cancer-related topics. More general clinical terminology can be found in SNOMED CT, the Systematized Nomenclature of Medicine Clinical Terms [53].

Although available independently, most of the mentioned resources are accessible via the UMLS Terminology Services [12] and BioPortal [13] (Table 3). Moreover, UMLS provides integration of these resources by linking the original identifiers to a single designated concept.

BioPortal is a web portal that provides access to one of the largest repositories of biomedical ontologies [60]. It provides a uniform mechanism to access biomedical ontologies and terminologies provided in different representation formats, including OBO and OWL. The associated web services provide programmatic access to these ontologies [61], which allows for their easy integration into the TM framework. Tables 4 and 5 provide a list of ontologies from BioPortal that are directly or indirectly related to cancer. In addition, one can define and share their own cancer-specific ontology on BioPortal and take advantage of its web services.

Other resources that can be used as a guideline as to how to structure and report cancer information extracted from free text are listed in Table 6. The College of American Pathologist developed cancer checklists that prescribe collection of all critical elements that should be reported for cancer specimens [4]. These checklists provide templates for the types of information a TM system should aim to extract from free-text medical reports. Specifically, in the breast cancer domain BI-RADS defines a hierarchy of terms to describe findings in mammograms together with mammography assessment categories [9]. Classification of findings in free-text medical reports requires more complex text analysis as well as domain-specific knowledge. In general, tumours are classified using TNM (Tumour-Node-Metastases) classification of malignant tumours, an internationally agreed-upon standard to describe and categorise cancer stages and progression [62]. It is based on the extent of the primary tumour (T), spread to nearby lymph nodes (N) and distant metastasis (M). TNM guidelines are cancer-specific.

In addition to formal knowledge resources, public web sites may be helpful to TM developers to familiarise with the domain. Table 7 provides a list of credible web sites.

5. Text data processing

In this section we focus on specific text processing tasks together with a review of methods and techniques used to solve them in the cancer domain. We differentiate between four major NLP tasks: named entity recognition (NER), information extraction (IE), text classification and information retrieval (IR). In order to assess the feasibility and compare different approaches, we first describe how the systems supporting these NLP tasks can be evaluated.

5.1. *Evaluation*

Most NLP tasks can be viewed as classification problems in which, given an instance, the system predicts its class label. For instance, an NER system in effect labels a phrase as a named entity. Similarly, an IR system will classify a document as relevant or irrelevant with respect to the user's information need. Various measures can be used to evaluate classification performance based on a confusion matrix, which contains information about actual (or known) labels and those predicted automatically by the system. Given a binary classification problem, there are four possible outcomes

Table 3 – Examples of cancer-relevant vocabularies and classification systems.

Vocabulary	Body	URL	UMLS ID	BioPortal ID	Study
CPT	AMA	http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page?	CPT	1504	[20,52]
ICD-9	WHO				[20,26,34,45]
ICD-9-CM	WHO		ICD9CM	1101	[23,52,54]
ICD-10	WHO	http://www.who.int/classifications/icd/	ICD10	1516	[45]
ICD-10-CM	WHO		ICD10CM	1553	–
ICD-O-3	WHO	http://www.who.int/classifications/icd/adaptations/oncology/en/index.html			[23,24,55,56]
MedDRA	ICH	http://www.meddra.org/	MDR	1422	[39]
MeSH	NLM	http://www.nlm.nih.gov/mesh/	MSH	1351	[34,57–59]
NCI Thesaurus	NCI	http://ncit.nci.nih.gov/	NCI	1032	[60]
SNOMED CT	IHTSDO	http://www.ihtsdo.org/snomed-ct/	SNOMEDCT	1353	[21,26,27,30,32,45]

Abbreviations: AMA, American Medical Association; ICH, International Conference on Harmonisation; IHTSDO, International Health Terminology Standards Development Organisation; NCI, National Cancer Institute; NLM, National Library of Medicine; WHO, World Health Organisation.

Table 4 – Examples of cancer-specific ontologies.

Name	Description	BioPortal ID
Breast Cancer Grading Ontology	Assigns a grade to a tumour starting from the three criteria of the next generation sequencing.	1304
Cancer Chemoprevention Ontology	Describes and semantically interconnect the different paradigms of the cancer chemoprevention domain.	3030
Cancer Research and Management ACGT Master Ontology	Represent the domain of cancer research and management in a computationally tractable manner.	1130
Neomark Oral Cancer Ontology	Describes the medical information necessary for early detection of the oral cancer reoccurrence.	1686
Upper-Level Cancer Ontology	Provides an upper-level ontology for cancer.	3178
NanoParticle Ontology	Represents the basic knowledge of physical, chemical and functional characteristics of nanotechnology as used in cancer diagnosis and therapy.	1083

Table 5 – Examples of cancer-related ontologies.

Name	Description	UMLS ID	BioPortal ID
Radiology Lexicon	A controlled terminology for radiology – a single unified source of radiology terms for radiology practice, education and research.	Not available	1057
WHO Adverse Reaction Terminology	An open-ended terminology for coding of adverse reaction terms.	WHO	1354
Physician Data Query	A part of NCI's comprehensive cancer information database, which contains expert summaries on a wide range of cancer topics, a listing of some 30,000 cancer clinical trials from around the world, a directory of genetics services professionals, the NCI Dictionary of Cancer Terms, and the NCI Drug Dictionary.	PDQ	1349
NCI SEER ICD Neoplasm Code Mappings	NCI Surveillance, Epidemiology and End Results (SEER) conversions between ICD-9-CM and ICD-10 neoplasm codes.	NCISEER	N/A

Table 6 – Examples of cancer reporting systems.

System	Body	URL	Study
CAP Cancer Protocols	CAP	http://www.cap.org/	[32,64]
TNM	UICC	http://www.uicc.org/resources/tnm	[30,32,65]
BI-RADS	ACR	http://www.acr.org/Quality-Safety/Resources/BIRADS	[14,17–19]

Abbreviations: ACR, American College of Radiology; CAP, College of American Pathologists; UICC, Union for International Cancer Control.

Table 7 – Examples of cancer-related information on the web.

Web site	URL
Cancer Research UK	http://www.cancerresearchuk.org
Centres for Disease Control and Prevention	http://www.cdc.gov/DiseasesConditions/az/c.html
MedicineNet.com	http://www.medicinenet.com/cancer/focus.htm
MedlinePlus	http://www.nlm.nih.gov/medlineplus/cancers.html
National Cancer Institute	http://www.cancer.gov/
NHS Cancer Screening Programmes	http://www.cancerscreening.nhs.uk/
NHS Choices	http://www.nhs.uk/conditions/Cancer
WebMD	http://www.webmd.com/cancer/
World Cancer Research Fund	http://www.wcrf-uk.org/
World Health Organization	http://www.who.int/cancer/en/

of a single prediction: both positive and negative predictions can be either true or false with respect to the actual labels. A *true positive* (TP) is a correct positive prediction, whereas a *true negative* (TN) is a correct negative prediction. Based on two types of errors that may be made by the system, we differentiate between a *false positive* (FP), which is an incorrect positive prediction, and a *false negative* (FN), which is an incorrect negative prediction. In statistics, these errors are referred to as *type I error* and *type II error* respectively. In case of NER and IE, the comparison between manually labelled entities and automatically predicted ones can be based on exact matching, in which the compared entities must match exactly in order to be considered correct, and/or partial matching, in which the compared entities only need to overlap.

Two measures typically used to evaluate NLP systems are precision and recall. *Precision* (P) calculated as TP/(TP + FP) represents the percentage of positive predictions that are correct. *Recall* (R) calculated as TP/(TP + FN) represents the percentage of positive instances that were predicted as positive and corresponds to the ability to find positive instances. Maximising these measures is based on optimising two different aspects of classification performance. Reducing the number of type I errors will increase precision, while reducing the number of type II errors will increase recall. However, precision and recall are naturally opposed in the sense that increasing precision will typically lead to decreasing recall and vice versa. Therefore, systems will often be compared on how well they balance precision and recall, which can be estimated using F_β *measure* calculated as $(1+\beta^2)\cdot P\cdot R/(\beta^2\cdot P+R)$. The parameter β is a non-negative real number used to weigh precision against recall in terms of relative importance. The *F1 measure* (or simply *F measure*), which gives no preference to either precision or recall, is usually reported.

While most NLP systems are evaluated using three measures (precision, recall and *F*-measure), the systems reviewed in this article have been developed with specific medical applications in mind. Therefore, the articles describing them often contain terminology more commonly used in medicine (diagnostic testing in particular). While we will be using terms such as precision and recall hereafter, here we will map them to their synonyms more readily understood in the medical community. Precision is also called *positive predictive value*, whereas recall is typically referred to as *sensitivity*. In addition, other measures motivated by medical applications are used to discuss the performance. *Specificity* calculated as TN/(TN + FP) represents the percentage of negative instances that were correctly predicted and corresponds to the ability to avoid false positives, thus complementing recall (i.e. sensitivity). *Negative predictive value* calculated as TN/(TN + FN) represents the percentage of negative predictions that are correct, thus complementing precision (i.e. positive predictive value). Finally, *accuracy* calculated as (TP + TN)/(TP + TN + FP + FN) represents the percentage of all predictions, either positive or negative, that are correct.

5.2. *Named entity recognition*

NER identifies and classifies words and phrases into predefined categories such as diseases, symptoms and drugs. NER is used mostly as a vehicle for feature extraction in order to support more complex NLP tasks such as IE, text classification and IR. Most approaches reviewed in this article relied on dictionary-based NER methods to recognise cancer types and gene names. The overwhelming majority used MetaMap to recognise concepts from UMLS [65], usually focusing on specific classes (or semantic types as they are called in the UMLS documentation) of concepts.

The biomedical domain exhibits high degree of terminological variation, which stems from the ability of a natural language to name a single entity in different ways. It has been estimated that approximately one third of term occurrences are variants [66]. The cancer domain is no exception, where various synonyms for each cancer type exist [67]. For example, *breast cancer* is alternatively referred to as *carcinoma of the breast* or *mammary neoplasm*. The variation phenomenon is further magnified by numerous synonyms of associated genes. For instance, alternative names of *breast cancer susceptibility gene 1* include BRCA1 and its orthographic variants BRCA-1 and BRCA 1 along IRIS, PSCP, BRCAI, BRCC1 or RNF53. Analysis of text data depends on the ability to automatically recognise all variants and normalise them by mapping them to a single entity they name. Resources that can support this process include dictionaries with good coverage of alternative names and explicitly link them together as well as software tools that can match them flexibly against text, thus allowing for unforeseen name variants. UMLS described earlier is a comprehensive dictionary of biomedical terms organised by their meaning. In addition, the National Library of Medicine (NLM) provides MetaMap, a software tool that can annotate UMLS terms in text and map them to the corresponding entity [65]. In the previous example it would recognise different ways of referring to breast cancer in text. Not surprisingly, MetaMap was used in many approaches to support recognition of named entities related to cancer (e.g. [5,30,32,59,68]).

Kang et al. [69] used MetaMap as a baseline and demonstrated how its performance on disease names (including cancers) can be further improved using a rule-based approach. Dictionary-based NER can only recognise entities if the names by which they are denoted in text are part of the dictionary. Kang et al. combined shallow parsing with a number of rules that adjust noun phrases and feed them back into the normalisation process to check whether they refer to known entities. Their error analysis highlighted problems associated with cancer type recognition often due to coordination. Their approach included coordination resolution in which part-of-speech and chunking information was used to reformat the coordination phrase such as *colorectal*, *endometrial* and *ovarian cancers* and recognise *ovarian cancers*, *colorectal cancers* and *endometrial cancers* as separate entities. Disease identification improved across all evaluation measures with additional processing over MetaMap alone: precision, recall and *F*-measure values rose from approximately 55%, 55% and 55–69%, 64% and 66% respectively.

These evaluation results were obtained on the Arizona Disease Corpus (AZDC), a set of 2856 PubMed abstracts annotated with disease names mapped to their UMLS identifiers [70]. However, most clinical reports are dictated and as such they naturally contain a higher number of grammatically incorrect sentences, misspellings, errors in phraseology, transcription errors, acronyms and abbreviations. Some of these abbreviations and acronyms tend to be highly idiosyncratic to a specific domain as well as local practice, and as such they cannot always be found in standardised dictionaries such as those included in UMLS [71]. Nassif et al. [18] illustrated this phenomenon with examples taken from mammography reports. They based their NER on the BI-RADS lexicon of terms that can be used to describe findings in mammograms (see knowledge resources above for more information). The BI-RADS lexicon differentiates between 43 distinct mammography features. These features are not uniformly described in mammography reports, i.e. radiologists often use different terms to refer to the same concept. Some of these synonyms are explicitly defined in the lexicon (e.g. *equal density* and *isodense*), but others need to be provided by experts (e.g. *oval* and *ovoid*). Additional expert knowledge can contribute to improve NER performance either by customising existing lexicons or by supplementing them with semantic grammars. Nassif et al. implemented a grammar consisting of rules that specify well-defined semantic patterns and the underlying BI-RADS categories into which they are mapped.

More often, customised or bespoke dictionaries are used to deal with term variability. For example, the CGMIM system for mining information about cancers and associated genes considers 21 major cancer types [55]. To deal with the various ways a given cancer might be referred to in the text (e.g. *breast cancer*, *breast tumour*, *breast carcinoma*, *mammary gland tumour*, *cancer of the breast*, etc.), they created a list of synonyms for each cancer type using the ICD-O and adding familiar lay terminology. The dictionary would be useful for NER across different cancer domains, but the source code does not seem to be available any longer.

Similarly, Xie et al. performed cancer name recognition by comparing the text with a cancer name dictionary also compiled from ICD-O [56]. The cancer classification has two axes: morphology (which describes the form and behaviour of the tumour) and topology (which describes the site of origin). A customised dictionary contains a collection of common names and an ICD-O code that consists of morphology and topology codes. Additionally, the dictionary includes abbreviations for some cancer names, e.g. *OSCC* was added as a synonym for *oral squamous cell carcinoma*. While the TM results are publicly available in a database, the dictionary is not readily accessible.

Fang et al. developed a cancer name entity recogniser as part of their MeInfoText system for mining gene methylation and cancer association information [72]. They combined a cancer dictionary and regular expression patterns. The dictionary of cancer names including their abbreviations was compiled from the previous version of the system [57] (originally extracted from the *Neoplasms by Site* (C04.588) section of the MeSH vocabulary) and public web sites (see in Knowledge resources above). The patterns used to identify cancer types were as follows: (1) <abbreviation>, (2) <tumour_site> <cancer-related_keyword>, and (3) (.+*oma* | *leukemia* | *leukaemia*). Abbreviations include acronyms such as *NPC* (*nasopharyngeal carcinoma*) and *CRC* (*colorectal cancer*). Cancer-related keywords form a specialised lexicon comprised of the following surface names: *cancer*, *tumor* (*tumour*), *neoplasm*, *carcinogenesis*, *tumorigenesis* and *metastasis*. With the exception of abbreviations, the matching strategies were case-insensitive.

As explained earlier using the work of Kang et al. [69], dictionary lookup approaches may not be sufficient for NER regardless of how comprehensive the underlying dictionary is. While Kang et al. used a rule-based approach to improve a dictionary-based NER, Jin et al. [73] relied on an ML approach to recognise clinical descriptions of malignancy presented in text. Their software, MTag, applies conditional random fields (CRFs) over syntactic and domain-specific features to extract strings of text corresponding to a clinician's or researcher's reference to cancer (malignancy type). Malignancy type is defined here as the full noun phrase encompassing a mention of a cancer subtype such that *neuroblastoma*, *localised neuroblastoma* and *primary extracranial neuroblastoma* are all considered to be distinct malignant type references. They considered identification of all variable descriptions of particular malignant types, such as *squamous cell carcinoma* (histological observation) or *lung cancer* (anatomical location), both of which are underspecified forms of *lung squamous cell carcinoma*. MTag was trained and tested on PubMed abstracts pertaining to cancer genomics recording precision, recall and *F*-measure values of 85%, 82% and 83% respectively. Given known issues associated with processing of clinical text, these values would be expected to drop, but since the system is based on ML, it could be re-trained on a clinical dataset in order to achieve similar performance in this sublanguage.

Obviously, other named entities apart from cancer types need to be recognised to support TM of related information, e.g. drugs, genes, treatments, findings, anatomical sites, etc. Dictionaries of these entities can be obtained from specialised databases and ontologies. For example, RxNorm (a resource related to UMLS) provides normalised names for clinical drugs and links them to many of the drug vocabularies commonly used in pharmacy management and drug interaction software [74]. It provides two services: a normalised naming system for

generic and branded drugs and a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems. In addition to the Gene Ontology (GO) being the standardised representation of gene and gene product attributes across databases [75], gene (and protein) name recognition has been a field of intense activity in the last decade (e.g. [76]) and there is a plethora of tools that can be readily applied. Information about anatomical entities is available in Foundational Model of Anatomy (FMA) [77], the most comprehensive ontology in this domain. It is accessible via BioPortal and is also integrated into UMLS as one of the source dictionaries. A preliminary investigation suggested that anatomical terminology is necessary for modelling cancer invasion [23].

For most other named entities, UMLS can be used as a vocabulary source. Indeed, many approaches used UMLS to obtain names of relevant entities by focusing on specific semantic types. UMLS organises terms into a hierarchy of over 130 semantic types, thus grouping named entities into broad categories, which can be used to focus on relevant portions of this comprehensive vocabulary [7]. For example, the caTIES system for coding and retrieval of surgical pathology reports and tissue specimens consists of 11 modules, one of which is a semantic-type filter, which removes concepts associated with unwanted semantic types following NER performed by MetaMap [59]. Heintzelman et al. compiled a dictionary of approximately 675,000 UMLS terms by considering 16 semantic types and utilised it in ClinREAD, a commercial healthcare-oriented rule-based NLP system, to mine pain information in patients with metastatic prostate cancer [37]. Harkema et al. used MetaMap to label entities of only three semantic types (*anatomical structure*, *neoplastic process* and *sign or symptom*) to support processing of colonoscopy reports [5]. Similarly, Schadow and McDonald used MetaMap to recognise entities from 20 semantic types to extract information from surgical pathology reports [68].

SNOMED CT, which is integrated into UMLS, can be navigated in a similar fashion using its own structure. For example, the SCENT system relies on a dictionary of approximately 1000 clinical entities related to morphology, anatomic site and procedural type [21]. To support reasoning about malignancy status based on information contained in pathology reports, this dictionary was enriched with additional information. Namely, the malignancy potential of each morphology entity was classified by up to four physicians with expert pathology or oncology knowledge. In general, dictionaries used in more specific approaches focusing on particular cancer type and associated clinical procedure are of manageable size and complexity and as such may be easily customised in lexical terms or enriched with additional semantics. For instance, Denny et al. compiled a list of colonoscopy terms, which consists of 26 UMLS concepts related to colonoscopy as well as five new terms added as local synonyms for existing UMLS concepts (*cscopy*, *C scope*, *C scopy*, *cscope* and *colonscopy*) to support processing of non-standardised terminology in medical notes [25].

In summary, MetaMap can be used to effectively identify UMLS terms in text and they can be restricted to semantic types that are relevant for specific domains and applications. However, when bespoke dictionaries need to be utilised, a different dictionary lookup method needs to be used. There is a wide choice of dictionary lookup tools available many of which are released as open source (e.g. LINNAEUS [78]), but before considering their use for clinical cancer applications, they need to be evaluated in this particular domain. ConceptMapper by IBM is an open source tool for classifying mentions in text based on standardised or proprietary terminologies and providing named entities as output [24]. It can be configured to use different search strategies or syntactic concepts. ConceptMapper provides similar functionality and performance to MetaMap without being tied solely to the UMLS dictionary. Any UIMA-compatible tokeniser can be used to pre-process both input text and the dictionary content. ConceptMapper then processes input text on a token-by-token basis, one span (e.g. sentence or noun phrase) at a time. The process of matching tokens in text against those in dictionaries can be customised with respect to case sensitive or insensitive matching, stemming, abbreviation expansions and spelling variants. The NER was evaluated in the colon cancer domain using related pathology reports obtained from the Mayo Clinic. Two types of entities were considered for evaluation: histological diagnoses and anatomical sites. Two separate dictionaries were initially compiled from ICD-O and later augmented with synonyms from the UMLS SPECIALIST Lexicon as well as common abbreviations, adjectival forms and commonly used shorthand expressions. Different configuration parameters were used in the experiments and the average values of precision, recall and *F*-measure were 91%, 88% and 89% for anatomical sites and 86%, 90% and 88% for histological diagnoses, which are relatively high for dictionary-based NER in clinical text.

5.3. *Information extraction*

IE selects specific facts about pre-specified types of entities and relationships of interest. For example, the 2009 i2b2 medication extraction challenge focused on the extraction of medication-related information including: medication name (m), dosage (do), mode (mo), frequency (f), duration (du) and reason (r) from hospital discharge summaries. In other words, free-text medical records needed to be converted into a structured form by filling a template (a data structure with the predefined slots) with the relevant information extracted (slot fillers). In this task, the sentence "*In the past two months, she had been taking Ativan of 3–4 mg q.d. for anxiety.*" should be converted automatically into a structured form as follows [79]:

m="ativan" || do="3–4 mg" || mo="nm" || f="q.d." || du="two months" || r="for anxiety"

where *nm* indicates that particular information was not mentioned. The given example illustrates the need to extract named entities (e.g. *ativan*) as well as quantitative information (e.g. dose), but also the extraction of relationships in order to link the extracted pieces of information to one another. In the previous section, we have discussed named entities of relevance in the cancer domain and how they can be recognised, which would be the first step in extracting cancer-relevant information. The second step requires modelling of other types of information (e.g. tumour size) and relating

it to the relevant entities (e.g. where multiple tumours are mentioned). In this section, we overview the types of cancer-related information that have been successfully extracted as well as techniques and approaches used to implement these IE tasks.

A lot of the extracted information will be specific to a given cancer type. For example, pathology reports related to prostate cancer will typically contain a Gleason score, which a pathologist uses to convey information about the severity of prostate cancer based on the appearance of cancer cells. The Gleason grade scale ranges from 1 to 5. Two Gleason values, primary and secondary, are determined and summed to obtain the final Gleason score, which ranges from 2 (1 + 1) to 10 (5 + 5) indicating the lowest and highest cancer aggression respectively. Napolitano et al. [36] provided examples of pathology reports and various ways in which Gleason score is referred to in text, e.g. "*Gleason grades 4 + 4, Gleason score 8*", "*Gleason score 8 – 10*", "*Gleason score is 7 (3 + 4)*", "*Gleeson grade 3. Total score 6.*", etc. The linear structure of this type of information makes it amenable to modelling with regular expressions. Indeed, Napolitano et al. achieved very high recall (98%) and precision (almost 100%) by taking such a simple approach. The candidate text lines were selected using the word *Gleason* as well as *Gleeson* as its common misspelling. The most common textual patterns used by the pathologist to record the Gleason grades (G1 and G2) and total score (S) were identified with some possible variants, e.g. "*Gleason score G1 + G2 = S*", "*Gleason score is G1 + G2 = S*", "*Gleason [. . .] (S) G1 + G2*", etc. The patterns were coded as regular expressions in Perl and made freely available at this URL: ftp://ftp.qub.ac.uk/pub/users/gnapolit/perl/. The potential benefits of this highly reliable approach to extracting cancer staging information are best illustrated with a fact that manual extraction of Gleason score from a test set of 915 reports was performed in about 30 person-hours as opposed to around 4 person-hours for coding and fine tuning of extraction rules followed by negligible processing time.

The same approach was also tested in the skin cancer domain. Pattern matching rules were implemented to extract two types of melanoma diagnostic indicators, Breslow depth and Clark level, from skin biopsy and excision biopsy reports. The uniform way in which this information is recorded by the pathologists enabled implementation of a rule-based method that reached 100% for both precision and recall on a set of 992 reports. These results demonstrated that this approach may increase the completeness of melanoma staging in Northern Ireland Cancer Registry by 32% and 18% for Breslow depth and Clark level respectively.

Similarly, Buckley et al. used commercial NLP software from ClearForest (a Thomson Reuters company) to demonstrate how a large body of free text medical information in breast pathology reports can be converted to a machine readable format using NLP [20]. Having an expert read and interpret each report is an effective but inefficient approach to unlocking the information in free text. It may be suitable in the day-to-day care of individual patients, but it is impractical on a large scale or when undertaking retrospective studies. The goal of this study was to automatically extract information about pathologic diagnoses from breast pathology reports, i.e. identify which specimens had evidence of diagnoses of interest: invasive ductal cancer (IDC), invasive lobular cancer (ILC), invasive cancer NOS, ductal carcinoma in situ (DCIS), severe atypical ductal hyperplasia (severe ADH), lobular carcinoma in situ (LCIS), atypical lobular hyperplasia (ALH), atypical ductal hyperplasia (ADH) and benign. To deal with the inherent linguistic and structural variability within free text, a lexicon was created for each diagnosis to hold a set of words and phrases that denote it. For example, the phrases such as *infiltrating ductal carcinoma*, *invasive cancer with ductal features*, *invasive cancer*, *ductal type*, etc. all went into the *invasive ductal cancer* lexicon. Some phrases were built into more than one lexicon, e.g. *invasive carcinoma with both ductal and lobular features* was placed in both IDC and ILC lexicons. A significant variability in the way that breast diagnoses can be expressed was noted. Excluding typographical and spacing errors, Buckley et al. identified 124 ways of saying invasive ductal cancer, 95 ways of saying invasive lobular cancer and almost 100 ways of referring to four other types of cancer considered. Lexicons were matched against text to perform NER. As the ultimate goal was to extract correct diagnoses and not only their mentions in text, further processing was needed. A diagnosis entity may be negated, e.g. a report may state that there was *no evidence of invasive carcinoma* or that *residual DCIS was not seen*. A pattern-matching approach was used to recognise negation (covering a total of 33 ways in which negation can be expressed) and remove negated diagnoses from further consideration. When multiple diagnoses were present, then the most significant diagnosis was determined based on the order of significance, e.g. IDC, ILC or invasive cancer NOS, would outweigh DCIS, which would outweigh severe ADH, which would outweigh LCIS, which would outweigh ALH, which would outweigh ADH, which would outweigh benign. The most significant diagnosis was proposed as the primary diagnosis, whereas the others were listed as secondary diagnoses. The evaluation on a set of 6711 pathology reports from three institutions resulted in sensitivity of 99%, specificity of 96%, positive predictive value of 99% and negative predictive value of 98%.

Yet another approach used regular expressions to extract four types of information including tissue type (e.g. liver), site modifier (e.g. right lobe), collection method (e.g. segmentectomy) and diagnosis (e.g. N4-NX-MX) from surgical pathology reports [68]. This study hypothesised that surgical pathology reports would be easier to parse than radiology reports, clinical notes or discharge summaries, because they often conform to a certain structure that can be exploited in IE. The results show that automatic coding of targeted information based on the UMLS identifiers was at least sufficient in 90% of cases.

Mamlin et al. demonstrated that coded information can also be reliably extracted from radiology reports [54]. They used a commercial system LifeCode [80] by A-Life Medical, Inc. now OPTUMInsight to code findings in cancer-related radiology reports, i.e. chest X-ray reports, based on ICD-9 and CPT. The precision and recall achieved on a set of 500 manually coded reports were 96% and 85% respectively. Their corpus is available on demand.

Structured information aggregated across different types of clinical records can support longitudinal analysis of cancer status, e.g. to predict survival in metastatic cancer. Heintzelman et al. used ClinREAD, a commercial healthcare-domain-oriented, rule-based NLP system by AeroText, to

extract pain status from all available paper, electronic, radiologic, radiation therapy and pathology records [37]. They focused on pain status in particular as it can be used as a predictor of survival in metastatic prostate cancer and an indicator of effectiveness of new therapies. Pain-related information included severity, body location and date (start and end). Pain status was classified on a four-tier pain scale: no pain, some pain, controlled pain and severe pain. With the average *F*-measure of 95% for pain mention detection and 81% for pain severity classification, this study successfully tested the feasibility of automatically tracking patient pain over time using NLP.

Early detection of cancer followed by appropriate treatment significantly improves survival rates. Radiologists may recommend additional imaging to improve cancer detection, but consistency between radiologists with regard to recommendation practices for similar patient and clinical attributes, increase or decrease in recommendation rates and patterns with evolving technology, type of recommended imaging techniques and time frames need to be examined in order to provide radiologists with specific guidelines for making appropriate recommendations. Dang et al. used LEXIMER, a commercial system by Nuance, to extract information on recommendation for additional imaging [81]. Rules were defined to extract imaging modalities (e.g. *MRI*, *MR angiography*, *MR spectroscopy*, etc.), recommendations (e.g. *recommend*, *suggest*, *follow up*, etc.) and time frame (e.g. *one year*, *six weeks*, etc.). Their approach reached accuracy of 93% for recommended imaging technique and 94% for time frames, thus demonstrating that accurate determination of recommended imaging techniques and time frames from radiology reports is possible with NLP.

Denny et al. implemented a method to extract similar types of information from EMR notes [25]. They specifically focused on colonoscopy as a diagnostic procedure for suspected colorectal cancer. Information about time reference and colonoscopy status was extracted using a set of linguistic and heuristic rules. Heuristics were developed for each status type relying on typical phrases as status indicators: scheduling (e.g. *referred for*, *ordered*), considering (e.g. would like to wait), discussion (e.g. *discussed*, *explained*, *recommended*), in need of (e.g. *due for*, *recommended*), receipt (e.g. *had*, *underwent*) and refusal (e.g. *refused*, *declined*). Timing references were extracted at recall of 91% and precision of 95%, colonoscopy status at recall 82% and precision of 95%, and colonoscopy completion at recall of 93% and precision of 95%. The system was later extended to extract information about other colorectal cancer screening methods in addition to colonoscopy: flexible sigmoidoscopy, faecal occult blood testing and double contrast barium enema [28]. The average recall and precision were 93% and 94%. The NLP method proved superior to the baseline defined as the associated billing codes explicitly recorded in the EMRs (44% recall and 83% precision) and as such is a useful adjunct to traditional methods to detect colorectal cancer screening status.

We conclude this section with an overview of MedTAS/P, a system that extracts complex information to obtain a structured cancer representation from free-text pathology reports [23]. The system uses a cascade of NLP techniques to populate Cancer Disease Knowledge Representation Model (CDKRM), which acts as a detailed template for extraction of cancer-related information. The main classes and their attributes in this model are as follows:

1. histology: mention, terminology code
2. anatomical site: mention, terminology code
3. grade: value, scale, type
4. dimension: extent, unit
5. date: day, month, year
6. gross description part: anatomical site, size
7. primary tumour: anatomical site, histology, size, grade
8. metastatic tumour: anatomical site, originating anatomical site, histology, size, grade
9. lymph node: anatomical site, histology, number of positive nodes, total number of nodes excised

MedTAS/P performs the usual linguistic pre-processing including tokenization, sentence discovery, part-of-speech tagging and shallow parsing. These tools needed to adapt not only to the medical domain, but to a specific sub-domain as well. In particular, the conventions and style of pathology reports needed to be taken into account. Errors such as tagging the word *nodes* as a verb instead of a noun in the context of *lymph nodes* propagate through the NLP pipeline and affect subsequent processing such as NER and relationship extraction. Therefore, the grammars for general English for the shallow parser were adapted to reflect the syntactic structure of pathology reports.

Linguistic pre-processing is followed by NER using the ICD-O and FMA vocabularies. Cancer-specific information (e.g. grade, stage, size, margin, date, tumour blocks, etc.) is extracted using a combination of rule-based and ML approaches. Finally, individual results were integrated by extracting relations between them and were used to populate the CDKRM.

The system was evaluated on a manually annotated set of pathology reports of patients diagnosed with colon cancer. The best results in terms of *F*-measure (97–100%) were achieved for instantiating classes in the CDKRM such as histologies or anatomical sites. A slightly lower *F*-measure (82–93%) was recorded for primary tumours or lymph nodes, which require the extraction of relations. The lowest *F*-measure (65%) was achieved for metastatic tumours, which was attributed to the small number of cases in the training. Indeed, other systems such as MEDTEX performed metastasis classification with very high accuracy (94%) [32]. Nonetheless, MedTAS/P provides an open-source platform which allows for further improvements and modification. The very detailed cancer representation model can capture cancer disease characteristics in a comparable and consistent fashion. It is extensible and modifiable thus allowing for the model to be re-used across different cancer domains. The system itself provides the mappings from free text to the model.

5.4. *Text classification*

IE converts free text data into structured information, which adds significant value to the data in terms of automated analyses that can then be performed over semantically typed and structured data. In layman terms, one could view IE as a

conversion of a Word document into an Excel spreadsheet, which makes complex statistical calculations only a click away. However, IE only applies to explicitly stated information. More value can be gained by inferring additional information that is not explicitly articulated in the original text. This can be achieved using text classification, which uses features extracted from text (e.g. using NER or more advanced IE) to map text (e.g. sentence, paragraph or most often a whole document) into one or more classes from a predefined scheme. Using breast cancer as an example, imagine classification of radiology reports into BI-RADS assessment categories: 0 (incomplete), 1 (negative), 2 (benign finding), 3 (probably benign), 4 (suspicious abnormality), 5 (highly suggestive of malignancy), and 6 (known biopsy – proven malignancy). IE will identify explicitly coded information, but text classification will aim to unlock implicitly coded information, making it amenable to further computational analysis where necessary.

In their early work, Burnside et al. achieved modest results in terms of precision (83.4 ± 5.3%) and recall (35.4 ± 5.6%) for classification of radiology reports into BI-RADS assessment categories [14]. They applied the linear least squares fit mapping algorithm [82] to very basic features based on word frequencies and terms from the BI-RADS lexicon. Later, they implemented a more advanced method that achieved much higher values across the three evaluation measures (97.7% precision, 95.5% recall and 97% *F*-measure) on the same dataset [18]. In fact, their automated method outperformed manual feature extraction (97.5% precision, 89.6% recall and 93% *F*-measure) at the 5% statistical significance level. Excellent performance can be attributed to a much richer feature set based on external knowledge sources such as UMLS as well as methods more tuned to this specific task. They developed a semantic grammar, which consisted of rules specifying well-defined semantic patterns and the underlying BI-RADS categories into which they are mapped.

The NLP system [18] was later utilised for binary classification of breast cancer as either invasive or DCIS based on information contained in mammography reports [19]. The features combined coded information (e.g. family breast cancer history, personal breast cancer history, prior surgery, palpable lump, screening vs. diagnostic, indication for exam, breast density, BI-RADS code left, BI-RADS code right, BI-RADS code combined and principal finding) with information extracted from free text using NLP (e.g. mass margin, calcification distribution, mammary lymph, etc.). They used inductive logic programming (ILP) to automatically build a classification model expressed as a set of logical if-then rules. If any rule applies, the model classifies a mammogram instance as a positive instance (i.e. invasive or DCIS, depending on the model). If no rule applies, then the model considers the mammogram instance to be negative (i.e. the alternative class). The benefit of this approach is that rules are induced automatically from the data, thus bypassing the knowledge elicitation bottleneck. Clinician tend to prefer rule-based systems because of their explanatory power as opposed to alternative ML approaches (e.g. support vector machines) whose "black box" models do not provide insight or explanation into the reasons for a particular classification. In this study, an open source ILP engine Aleph [83] was applied to age-stratified mammography reports to build age-specific classification models of invasive versus DCIS cancer occurrence across the strata. An example of a rule predicting invasive cancer in the younger cohort: *The mammogram has a palpable lump in a breast, its breast density is class 2, and its calcification distribution is not reported*. In general, ILP provided a number of interesting rules, some of which were previously unreported and worthy of further investigation.

SCENT has been recently developed as a system that combines the functionality of the commercial SAS software with regular expressions and hierarchical decision rules to classify pathology reports according to the malignancy status (benign, borderline, basaloid or malignant) and to classify malignancy as either primary or recurrent [21]. Malignancy status was inferred by considering a set of clinical concepts from SNOMED CT codes related to morphology, anatomic site and procedural type in a cascade of decision rules. The malignancy potential of each morphology concept was previously classified by up to four physicians with expert pathology or oncology knowledge. Morphology codes and anatomic sites consolidated into categories are used to differentiate between new primary and recurrent malignancies. Anatomic site classifications were consolidated into categories, e.g. the sternum and clavicle sites belong to the bone category, sites relating to regional disease spread (e.g. neck and groin) belong to the lymph nodes category, etc. The system performance was evaluated in two cancer domains. The results for breast cancer were as follows: sensitivity 93–100%, specificity 98–100%, positive predictive value 85–98% and negative predictive value 99–100%; the results in the prostate cancer domain were equally good: sensitivity 89–99%, specificity 98–100%, positive predictive value 89–99% and negative predictive value 98–100%. The source code should soon become freely available for non-commercial use and modification from http://www.kp-scalresearch.org/research/tools_scent.aspx.

Several other systems have been implemented with a goal of automatically classifying cancer stage as a richer and more informative account of disease extent. TNM classification of malignant tumours is an internationally agreed-upon standard to describe and categorise cancer stages and progression [62]. It is based on the extent of the primary tumour (T), spread to nearby lymph nodes (N) and distant metastasis (M):

Primary tumour (T)
- TX: primary tumour cannot be evaluated.
- T0: no evidence of primary tumour.
- Tis: carcinoma in situ.
- T1, T2, T3, T4: size and/or extent of the primary tumour.

Regional lymph nodes (N)
- NX: regional lymph nodes cannot be evaluated.
- N0: no regional lymph node involvement.
- N1, N2, N3: degree of regional lymph node involvement (number and location of lymph nodes).

Distant metastasis (M)
- MX: distant metastasis cannot be evaluated.
- M0: no distant metastasis.
- M1: distant metastasis is present.

Most cancer types have TNM classification and the types of tests used for staging will depend on the type. In general, information gathered from physical exams, laboratory tests, imaging studies, pathology reports and surgical reports will

be used to determine the stage of a cancer. Several studies attempted to automatically produce TNM classifications from pathology reports [27,30,32,64].

CSIS (Cancer Stage Interpretation System) classifies the lung cancer stage based on information from pathology reports across two dimensions of the TNM system, T and N, but not M [64]. The system used text classification techniques to train support vector machines (SVMs) to extract elements of stage listed in cancer staging guidelines (e.g. maximum tumour dimension, visceral pleural invasion, main bronchus invasion, chest wall invasion, etc.) and then combined them into a global decision. CSIS achieved accuracy of 74% for tumour staging and 87% for node staging.

MEDTEX achieved similar performance in terms of accuracy (72%) for tumour staging of lung cancer using a rule-based method [32]. The accuracy for node staging was lower (78%), but MEDTEX includes metastasis classification at high accuracy (94%). The method is based on extracting information specified in a relevant CAP electronic Cancer Checklist (see Knowledge resources above), which are subsequently used to compute the TNM stage following the logic based on the staging guidelines. The SNOMED CT encoded version of the CAP cancer checklist relating to lung cancer resections was used, which made a MetaMap a natural choice for finding relevant entities in free text reports.

Martinez et al. experimented with different types of ML (naïve Bayes, SVM, Bayesian network and random forest) to perform TNM classification of colorectal cancer based on information found in histopathology reports [27,30]. The best results were achieved with Bayesian methods. The highest *F*-measure values for tumour, node and metastasis classification at 64%, 68% and 74% respectively leave room for improvement, but do indicate that automatic TNM stage classification of colorectal cancer is feasible with appropriate feature engineering.

Staying in the colorectal cancer domain, Harkema et al. demonstrated that the information required for colonoscopy quality measures such as "If indication is chronic diarrhoea, obtain biopsy" is amenable to automatic extraction from free-text colonoscopy and pathology reports [5]. Their NLP system extracted the values of the 21 necessary variables (e.g. indication type, biopsy, etc.) with an average accuracy of 89% and average F-measure of 74%, which imply that the NLP-derived outcomes for these quality measures can be practically useful for quality reporting. Moreover, with further refinement and development the NLP system could be employed for routine quality measurement on a large scale. At the moment, the values of the target variables are established using generally simple rules, e.g. if a concept *colon cancer* with contextual properties *directionality = affirmed, temporality = historical*, and *experiencer = family member* has been identified within a report, then the variable *family history* is set to *present*, whereas temporal expressions and size measurements are parsed and interpreted with a set of regular expressions.

In another rule-based approach, Wagholikar et al. relied on NLP to implement a clinical decision support system (CDSS) for colonoscopy surveillance using information found in pathology reports, procedure notes and the indications for colonoscopy [29]. This work was based on the premises that the guidelines for colonoscopy surveillance were comprehensive to address all possible patient scenarios and that the relevant patient information in the free text reports could be accurately extracted with NLP. Guidelines for colonoscopy surveillance were converted into a flowchart, which was implemented as set of 172 if-then rules that represented 43 nodes and 88 edges from the flowchart. The text processor was based on dictionaries, word patterns, and hand-coded rules to map the word patterns to parameter values of interest. Additional knowledge was elicited from pathologists and physicians, and coded into rules in order to interpret implicit information in text documents, e.g. if the pathologist does not mention a finding of cancer, the referring physician can interpret that there was no finding of cancer. All rules were combined in a binary classifier that outputs a decision for colonoscopy. Manual evaluation revealed that in 45 of 53 cases, the recommendation of the CDSS matched with the initial recommendation of the gastroenterologist; in five cases the gastroenterologist retained her initial recommendation, as the CDSS recommendation was not optimal; in three cases the CDSS helped the gastroenterologist to revise her initial blinded recommendations. Overall, the results demonstrated NLP can be effectively coupled with clinical decision support to improve colonoscopy surveillance for prevention of colorectal cancer.

Similarly, the results of the iDiagnosis system can be used to recommend screening for those at higher risk of pancreatic cancer automatically detected from EMR data [34]. Knowledge extracted from PubMed and EMRs semi-automatically was used to inform a prediction model. Variables of the model fall into five categories (demographics, life style, symptoms, co-morbidities and lab test results) and were selected based on a literature review, the recommendations by clinical experts and risk factors previously identified by the authors. All 20 variables were available in EMRs, but half of them could only be found in narrative notes and were searched using the MedLEE system [84]. All variables were weighted based on their associations with pancreatic cancer mined from PubMed. Keywords were used to indicate the polarity of each co-occurrence: positive (e.g. *risk*, *link*, etc.), negative (e.g. *comparison*, *discrimination*, etc.) or neutral (e.g. *equal to*, *same to*, etc.). The ratio between positive and negative associations was treated as a statistical summary of the collective evidence for the association between a risk factor and pancreatic cancer. Prior probability for each variable, e.g. *P*(*alcohol or cigarette abuse* = true | *pancreatic cancer* = true), was calculated using the EMR data. The weights and prior probabilities were then incorporated into the Bayesian Network Inference (BNI) model, from which the posterior probabilities, e.g. *P*(*pancreatic cancer* = true | *alcohol or cigarette abuse* = true), can be calculated and used for disease prediction. With sensitivity, specificity and accuracy of 85%, the BNI method significantly outperformed an SVM used as a baseline, which achieved 77% sensitivity, 40% specificity and 53% accuracy. This indicates that the choice of classification method largely depends on the dimensionality of feature space. SVM methods excel primarily in a high-dimensional feature space, whereas the BNI model performs well in a low-dimensional feature space.

A range of ML classifiers were studied for the classification of cancer-related death certificates: SVM, multinomial naïve Bayes, C4.5 and adaptive boosting [45]. Numerous features were used including stemmed words, bi-grams and SNOMED

CT terms, which were extracted with the MEDTEX toolkit [32]. An SVM classifier achieved the best *F*-measure of almost 99% when evaluated on a set of 5000 free text death certificates. The SVM classifier used with different features accounted for the top 18 of 40 evaluated runs and had the lowest variance, making it the most robust classifier for the task. The selection of features significantly influenced the performance of the classifiers. Stemmed tokens arose as the single most important feature set among those considered. The study found that SNOMED CT features provided consistent increments in classification robustness if used along with stemmed tokens.

Stemmed tokens combined with their TF-IDF score (a measure typically used in IR) also proved to be good features for classification of risk types of each human papillomavirus (HPV) based on their textual explanation in the HPV Sequence Database [41]. The database contains a complete list of papillomavirus types and hosts and the records for each unique papillomavirus [85]. Human papillomavirus infection is known as the main factor for cervical cancer, a leading cause of cancer deaths in women. With more than 100 types of HPVs it is critical to discriminate between those related with cervical cancer and those that are not. Each HPV type was represented as a vector whose elements were TF-IDF of 1434 stemmed tokens remaining after removing stop words. Three ML approaches were used: naïve Bayes, adaptive boosting and its variant AdaCost, a misclassification cost-sensitive boosting method. The latter performed best with accuracy of 93% and *F*-measure of 86%, thus implying that most high-risk HPVs were identified.

5.5. *Information retrieval*

IR can be viewed as a classification problem in which each document is classified as either relevant or irrelevant for the user's information need expressed with a search query. It is the necessary first step in gathering relevant data from a larger collection such as EMRs. Translational research requires detailed clinical information such as disease stage, disease severity and response to treatment to identify relevant cases and perform correlative studies. However, most clinical outcome information is stored as free text rather than coded, structured data. The caTIES system was implemented to make unstructured clinical information in surgical pathology reports more readily accessible for translation research [59]. It creates a text search engine index for fast access to documents based on text characteristics and conceptual codes and also maintains an ancestor index that associates NCI Thesaurus concepts with their ancestry. caTIES supports both query by text and query by concept. Queries can be constrained by demographic variables such as age and gender. Boolean operators AND, OR and NOT can be used to combine search terms and constraints. Additionally, temporal queries based on the timing of diagnostic reports are allowed, e.g. "*Find all females who had lobular carcinoma in situ, followed by mastectomy within 1 year*."

The precision of the caTIES system was 94–96% for simple and moderately complex queries, while it dropped down to 88% for more complex temporal queries. Most common errors referred to retrieval of documents in which the search concept was erroneously coded by the system because a substring of the more complex concept was recognised by MetaMap, e.g. report for *post-mastectomy scar* retrieved for query *mastectomy*. These errors often occur because the more complex concepts are post-coordinated concepts and are not represented in the dictionary. This emphasises the importance of NER performance in subsequent text processing. The previous section on NER provides descriptions of some approaches to dealing with these problems.

Another common source of errors was an incorrect clinical diagnosis, where specimen was labelled with a clinical diagnosis, which was subsequently corrected by pathological examination. Similarly, diagnostic uncertainty (e.g. *cannot exclude*) was a common problem in retrieving clinical documents. Other error categories observed include initials incorrectly coded as abbreviations (e.g. report with initials *HL* retrieved for query *Hodgkin's lymphoma*), concepts identified in the report that are in fact historical (e.g. report describing *previous history of renal cell carcinoma* retrieved for query *renal cell carcinoma*), conceptual relationships not properly scoped (e.g. report containing *prostate cancer without perineural invasion, and urothelial cancer with perineural invasion* is returned for query *prostate cancer with perineural invasion*), errors in negation detection (e.g. *neither prostatic intraepithelial carcinoma nor carcinoma is seen* is returned for query containing *prostatic intraepithelial carcinoma*). Most of the observed errors could be eliminated by limiting search to specific report sections and by extending the negation detection to account for uncertainty. The caTIES system is freely available under an open source licence.

Evaluation of another clinical retrieval system, ARC [26], showed variation across three cancer domains: colorectal, prostate and lung cancer. The data used for evaluation included imaging reports (X-rays, CT scans and MRI) consistent with lung cancer, pathology reports consistent with colorectal cancer and pathology reports consistent with prostate cancer. Recall, precision and *F*-measure were 90%, 92% and 89% for colorectal cancer, 97%, 95% and 94% for prostate cancer and significantly lower at 76%, 80% and 75% for lung cancer. This is not necessarily the reflection of the language variability across the three domains, but rather the type of information contained in different types of reports. The pathology report is the primary document for recording a diagnosis of prostate cancer and colorectal. However, lung cancer diagnosis is usually determined by a combination of imaging studies, biopsies and laboratory results. Despite the variability in the results, this study demonstrated that the ARC system with no custom software or rules development is sufficiently generalisable across different cancer domains. The ARC system is also freely available under an open source licence.

We will end this section and the review of different NLP tasks in the cancer domain by relating it back to patients. Bader and Theofanos analysed cancer-related queries on Ask.com, a search engine which allows users to create queries using whole phrases and sentences of any length rather than just key words [86]. Over 78% of sampled queries referred to 14 cancer types. The most-common cancer types mentioned in queries were digestive/gastrointestinal/bowel (15%), breast (12%), skin (11%) and genitourinary (11%). Queries were sorted into cancer-related categories including general information,

symptoms, diagnosis and testing, treatment, statistics, definition and cause/risk/link. Additional categories of queries about specific cancer types varied. This study highlighted the specific needs of patients and general-public users, what they really want to know about cancer, how they phrase their questions and how much detail they use. These results obtained with NLP can help healthcare providers improve the way in which they communicate cancer-related information to patients.

6. Text mining systems

A table in the Supplementary material provides a summary of the systems described in this section focusing on particular NLP tasks: named entity recognition, information extraction, text classification and information retrieval. Here we provide a more detailed overview of two more generic NLP systems that have been developed for and/or tested in the cancer domain.

6.1. *MedLEE*

The goal of the MedLEE (Medical Language Extraction and Encoding) system, developed at Columbia University in collaboration with the City University of New York, is to extract, structure and encode clinical information in free text patient reports so that it can be further exploited by subsequent automated processes. Originally, MedLEE was designed to process radiological reports of the chest to detect patients suspicious for tuberculosis [84]. MedLEE has since been extended to cover all of radiology and also pathology, echocardiology, electrocardiography and hospital discharge summaries [87]. Being tuned for processing radiology and pathology reports makes MedLEE directly relevant for the cancer domain. In particular, MedLEE was used to mine breast cancer information from surgical pathology reports [16] as well as mammogram reports [88]. It was also applied to process 889,921 chest radiographic reports in order to extract information about 24 medical conditions including lung cancer [31].

MedLEE consists of the following modules: pre-processor, parser, phrase regularisation and encoding. Pre-processor segments the text into sections, paragraphs, sentences and words. It then performs NER by dictionary lookup, handles abbreviations using a mapping table, and performs some word sense disambiguation based on contextual rules. The initial tuning of MedLEE for a specific cancer domain would apply to these elements, i.e. lexicons, tables and contextual rules.

The parser determines the structure of each sentence using a grammar that consists of syntactic and semantic rules. Following the parsing stage, the phrases are regularised in order to normalise the representation of their meaning. One aspect of phrase regularisation involves using domain knowledge to add information to the output that is implicit in the domain. For example, *infarct* implies *myocardial infarction* in cardiology reports, but could refer to another body location in other types of reports (e.g. *pulmonary infarction*). The domain knowledge is specified in a table created manually using domain expertise, so it would require knowledge elicitation in order to model different cancer domains. Encoding uses another table to add codes (e.g. UMLS) to words and regularised phrases. This means that tables can be created to enable the use of different coding systems. The section on Knowledge sources lists coding systems relevant to cancer that can be utilised in this module.

We will use two case studies to illustrate how well MedLEE can perform in cancer domains. Hripcsak et al. performed internal and external validation to investigate the accuracy of NLP for translating chest radiographic narrative reports into a large database [31]. MedLEE was used to code 889,921 reports on 251,186 patients. Using a set of 150 manually coded reports as a gold standard, sensitivity of 81% and specificity of 99% were reported. A total of 24 clinical conditions (diseases, abnormalities and clinical states) were the subject of this study. We believe that focusing on lung cancer alone would allow for finer tuning of the underlying lexical and domain-specific knowledge, which would be reflected in better sensitivity.

In another study MedLEE was extended to better deal with breast cancer information found in surgical pathology reports [16]. Targeted information included procedure name, number of positive lymph nodes, expression of oestrogen receptors, progesterone receptors and Her-2/Neu, nuclear grade, ploidy, DNA index, quantitative S-Phase, qualitative S-Phase, G2-M and proliferation Index. A relatively small number of manually annotated documents were used for development (20) and testing (50). High values for sensitivity (91%) and precision (92%) were recorded. However, performance was much better for the tabular findings (sensitivity of 96% and precision of 95%) than for narrative findings (sensitivity of 86% and precision of 88%).

6.2. *cTAKES*

cTAKES (clinical Text Analysis and Knowledge Extraction System), a generic NLP system developed at the Mayo Clinic, is tailored to the clinical domain and can add rich linguistic and semantic annotations to the narrative found in EMRs [89]. It has been designed to be scalable, comprehensive, modular, extensible and robust to meet the rigours of clinical research. cTAKES consists of the following NLP modules: sentence boundary detector, tokenizer, normalizer, POS tagger, shallow parser and NER annotator (including status and negation annotators). The performance of individual components was evaluated on a clinical corpus sampled from the Mayo Clinic EMRs. Sentence boundary detection was performed with an accuracy of 95%. Tokenizer achieved the same accuracy with part-of-speech tagger accuracy only slightly lower at 94%. The shallow parser achieved an F-measure of 93%. NER performance (F-measure of 72% for exact matching and 82% for non-exact matching) could be improved with richer dictionaries and additional post-processing. Negation and status attributes of named entities were extracted with an F-measure of 96% and 94% respectively. A global system evaluation of cTAKES for patient cohort identification for 25 clinical research studies has been conducted. The system was also used for treatment classification for a pharmacogenomics breast cancer treatment study. The cTAKES named entity attributes are similar to those in MedLEE. Unlike MedLEE, the cTAKES currently does not assert relationships between a disease/disorder, sign/symptom or procedure and an anatomical

site. However, MedLEE is a patented system and is undergoing commercialization via Health Fidelity, Inc., who were given exclusive rights to its portfolio of patents, software code and trademarks. In contrast, cTAKES has been released under a free software licence and thus can be readily used to support NLP applications for cancer but can also be further modified to better address specific tasks in this domain. It is available for download from http://ctakes.apache.org/.

7. Discussion and conclusions

In this article, we have reviewed current TM approaches with clinical applications in the cancer domain. The review highlighted a strong bias towards symbolic methods, e.g. NER largely based on dictionary lookup and IE relying on pattern matching. While these can be regarded as more conservative approaches to TM, they nonetheless deliver good performance. The *F*-measure of NER ranges between 80% and 90% (e.g. [73]), while that of IE for simple tasks is in the high 90s (e.g. [36]). This can be explained by the relative stability of the clinical sublanguage in this particular domain in comparison to biology, where dictionaries quickly become out of date, so ML approaches such as CRFs are more commonly used for NER (e.g. [90]). The comparatively slower dynamics of the clinical cancer domain allows for dictionaries to be updated manually when new concepts such as drugs, diagnostic tests and related genes are identified. However, the fact that clinical concept names vary considerably due to idiosyncrasies of the clinical sublanguage such as non-standard abbreviations (e.g. *cscopy*, *C scopy* are used as synonyms of *colonoscopy*) can be attributed to NER relying on standard dictionaries (e.g. UMLS) and tools (e.g. MetaMap) not performing better than 90% in terms of *F*-measure.

The relatively limited domain scope still allows for manual extension of dictionaries with non-standard terminology (e.g. [21]), but the problems associated with the clinical sublanguage such as high degree of spelling and grammatical errors calls for a more general and robust approach to NER. To account for syntactic variation, Kang et al. [69] demonstrated how a rule-based approach mapping between the shallow parses and MetaMap outputs can improve the results of NER. Our investigation into automatic term recognition revealed that most syntactic, morphological and orthographic variations can be identified by simply ignoring the internal phrase structure in a bag-of-words approach and using phonetic and edit distance algorithms to map between individual tokens [91]. We, therefore, suggest that such an approach can be used to improve the performance of dictionary-based NER in order to effectively deal with both syntactic variation and spelling errors.

Nonetheless, there remains a need for a comprehensive cancer ontology that would include rich terminologies for each cancer type as well as relationships to other relevant concepts (e.g. screening and treatment methods, synoptic reporting formats, checklists, etc.). The ontology should be an open community effort allowing seamless integration of internally developed terminologies, which currently are either not shared at all (e.g. [56]) or remain confined to supplementary materials (e.g. [25]) in non-standardised formats. The time saved on collection of vocabularies and construction of bespoke ontologies (e.g. [22]) and representation models (e.g. [23]) would allow more resources for development of the actual TM methods. Sharing the ontology in community portals comes with the benefit of readily available web services (e.g. [61,92]) that would automatically provide programmatic access to the ontology, which would allow easy integration into the TM framework [93].

Most IE systems reviewed here are rule-based. For simple types of information, such methods are fast and easy to develop and they perform extremely well (e.g. [36]). However, for more complex and variable types of information, ML may be necessary. A range of synoptic reporting formats available for different types of cancer [4] clearly indicate what information should be targeted by future IE systems. Synoptic formats can act as IE templates. Where synoptic reports are already used in clinical practice, their content can be used to train the IE systems.

The extent to which ML will be incorporated into IE system will undoubtedly be limited by strong preference for rule-based systems over "black box" ML models in clinical practice. Some ML approaches will be more acceptable than others. For example, approaches such as ILP can be used to induce rules, thus effectively addressing the knowledge elicitation bottleneck associated with rule-based systems while still benefiting from the explanatory power of rules (e.g. [18]). In many cases, the reporting guidelines such as those for TNM staging will readily provide the classification rules, hence the need for ML may be confined to extracting the relevant information not necessarily classification itself.

Narrative reports such as pathology and radiology reports convey valuable diagnostic information that is predictive of the prognosis and biological behaviour of a disease process [63]. A range of studies have proven the feasibility of NLP for structuring free text reports (e.g. [16,20,23,27,94]). Evaluations of NLP systems suggest that they can provide accurate data on service provision and patient clinical status [95]. In some cases, NLP results were superior to formally coded information in EMRs (e.g. [25]), which opens a range of possible applications in clinical practice. Indeed, many NLP approaches have been developed with direct clinical applications for cancer monitoring and prevention in mind. For example, two studies have shown that NLP can be coupled with clinical decision support in order to improve colorectal cancer screening rates [25,29]. Unfortunately, the actual procedures for colorectal cancer screening are often inadequate and vary widely among physicians, but NLP can support routine measurement of colonoscopy quality [5]. Similarly, NLP was applied to radiology reports to investigate additional imaging recommendation practice as this task requires careful balance between healthcare cost and accuracy and timeliness of cancer diagnosis [81]. Most often, NLP was used for predictive modelling of cancer that can be used to inform clinical decision making. Examples include prediction of invasive cancer across different age cohorts [19], risk of malignancy of breast cancer [17], risk of pancreatic cancer [34], lung cancer stage [32], survival in metastatic prostate cancer [37], etc. Most of these NLP systems have been developed and used locally due to strict legal regulations, which make healthcare institutions reluctant to share free-text medical reports even when all reasonable measures to protect privacy and

Summary points
What was already known before this study:

- Narrative reports such as pathology or radiology reports convey valuable diagnostic information that can be used for predictive modelling of cancer.
- Individual studies have proven the feasibility of specific NLP techniques for extracting structured information from free text reports.

What this study has added to our knowledge:

- An overview of available knowledge sources that can support semantic interpretation of clinical narratives related to cancer.
- An overview of available methods and techniques used for TM of clinical narratives related to cancer.
- The finding that ML approaches have not been fully explored in TM of cancer-related information.

confidentiality are taken (e.g. de-identification and data use agreements) [51]. More widely, this issue remains the main bottleneck for progress in healthcare applications of NLP.

Authors' contribution

JAK and GN conceived the overall study and scoped the review. IS conducted the literature search and drafted the manuscript. JL provided medical comments on the paper. All authors read and approved the final manuscript.

Conflict of interest

The authors declare there is no conflict of interests.

Acknowledgements

This work was partly funded by The Christie NHS Foundation Trust. GN acknowledges support from the Health e-Research Centre (HeRC) and Serbian Ministry of Education and Science (projects III44006; III47003). The authors wish to acknowledge the following members of the Clinical Outcomes Unit at The Christie NHS Foundation Trust: Matt Barker-Hewitt, Tom Liptrot, Catherine O'Hara and Ben Wilson.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009.

REFERENCES

[1] Cancer Research UK, UK Cancer Incidence (2010) by Country Summary, 2013 http://cancerresearchuk.org/cancer-info/cancerstats/
[2] Office for National Statistics, Deaths Registered in England and Wales, 2012, 2013 http://www.ons.gov.uk/ons/rel/vsob1/death-reg-sum-tables/2012/sb-deaths-first-release–.html
[3] College of American Pathologists. http://www.cap.org/, 2013.
[4] Centers for Disease Control and Prevention, Cancer Data and Statistics Tools, 2013 http://www.cdc.gov/cancer/npcr/tools.htm
[5] H. Harkema, W.W. Chapman, M. Saul, E.S. Dellon, R.E. Schoen, A. Mehrotra, Developing a natural language processing application for measuring the quality of colonoscopy procedures, J. Am. Med. Inform. Assoc. 18 (2011) i150–i156.
[6] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, et al., Biomedical text mining and its applications in cancer research, J. Biomed. Inform. 46 (2013) 200–211.
[7] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) D267–D270.
[8] R.A. Baasiri, S.R. Glasser, D.L. Steffen, D.A. Wheeler, The Breast Cancer Gene Database: a collaborative information resource, Oncogene 18 (1999) 7958–7965.
[9] E.S. Burnside, E.A. Sickles, L.W. Bassett, D.L. Rubin, C.H. Lee, D.M. Ikeda, et al., The ACR BI-RADS® experience: learning from history, J. Am. Coll. Radiol. 6 (2009) 851–860.
[10] National Cancer Institute, NCI Thesaurus, 2013 http://ncit.nci.nih.gov/
[11] US National Library of Medicine, Medical Subject Headings (MeSH), 2013 http://www.nlm.nih.gov/mesh/
[12] US National Library of Medicine, UMLS Terminology Services, 2013 https://uts.nlm.nih.gov/
[13] National Center for Biomedical Ontology, BioPortal, 2013 http://bioportal.bioontology.org/
[14] B. Burnside, H. Strasberg, D. Rubin, Automated indexing of mammography reports using linear least squares fit, in: 14th International Congress and Exhibition on Computer Assisted Radiology and Surgery, 2000, pp. 449–454.
[15] C. Blake, W. Pratt, Better rules, fewer features: a semantic approach to selecting features from text, in: IEEE International Conference on Data Mining, San Jose, CA, 2001, pp. 59–66.
[16] H. Xu, K. Anderson, V.R. Grann, C. Friedman, Facilitating cancer research using natural language processing of pathology reports, Stud. Health Technol. Inform. 107 (2004) 565–572.
[17] E.S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M.J. Lindstrom, B.M. Geller, et al., Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings, Radiology 251 (2009) 663–672.
[18] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, D. Page, Information extraction for clinical data mining: a mammography case study, in: IEEE International Conference on Data Mining, Miami, FL, 2009, pp. 37–42.
[19] H. Nassif, D. Page, M. Ayvaci, J. Shavlik, E.S. Burnside, Uncovering age-specific invasive and DCIS breast cancer rules using inductive logic programming, in: T. Veinot (Ed.), 1st ACM International Health Informatics Symposium, Arlington, VA, 2010, pp. 76–82.
[20] J.M. Buckley, S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A.K. Belli, et al., The feasibility of using natural language processing to extract clinical information from breast pathology reports, J. Pathol. Inform. 3 (2012) 23.
[21] J.A. Strauss, C.R. Chao, M.L. Kwan, S.A. Ahmed, J.E. Schottinger, V.P. Quinn, Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm, J. Am. Med. Inform. Assoc. 20 (2013) 349–355.

[22] J. Polpinij, A. Miller, Ontology-based text analysis approach to retrieve oncology documents from PubMed relevant to cervical cancer in clinical trials, in: P. Perner (Ed.), ICDM Workshop on Advances in Data Mining, IBaI Publishing, Leipzig, Germany, 2010.
[23] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K.S.J. Cooper, et al., Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, J. Biomed. Inform. 42 (2009) 937–949.
[24] M. Tanenblatt, A. Coden, I. Sominsky, The ConceptMapper approach to named entity recognition, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al (Eds.), Language Resources and Evaluation, European Language Resources Association, Malta, 2010, pp. 546–551.
[25] J.C. Denny, J.F. Peterson, N.N. Choma, H. Xu, R.A. Miller, L. Bastarache, et al., Extracting timing and status descriptors for colonoscopy testing from electronic medical records, J. Am. Med. Inform. Assoc. 17 (2010) 383–388.
[26] L.W. D'Avolio, T.M. Nguyen, W.R. Farwell, Y. Chen, F. Fitzmeyer, O.M. Harris, et al., Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC), J. Am. Med. Inform. Assoc. 17 (2010) 375–382.
[27] D. Martinez, Y. Li, Information extraction from pathology reports in a hospital setting, in: 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 2011, pp. 1877–1882.
[28] J.C. Denny, N.N. Choma, J.F. Peterson, R.A. Miller, L. Bastarache, M. Li, et al., Natural language processing improves identification of colorectal cancer testing in the electronic medical record, Med. Decis. Making 32 (2012) 188–197.
[29] K. Wagholikar, S. Sohn, S. Wu, V. Kaggal, S. Buehler, R. Greenes, et al., Clinical decision support for colonoscopy surveillance using natural language processing, in: IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, San Diego, CA, 2012, pp. 12–21.
[30] D. Martinez, L. Cavedon, G. Pitson, Stability of text mining techniques for identifying cancer staging, in: H. Suominen (Ed.), 4th International Workshop on Health Document Text Mining and Information Analysis, Sydney, Australia, 2013.
[31] G. Hripcsak, J.H. Austin, P.O. Alderson, C. Friedman, Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports, Radiology 224 (2002) 157–163.
[32] A.N. Nguyen, M.J. Lawley, D.P. Hansen, R.V. Bowman, B.E. Clarke, E.E. Duhig, et al., Symbolic rule-based classification of lung cancer stages from free-text pathology reports, J. Am. Med. Inform. Assoc. 17 (2010) 440–445.
[33] A.R. Tate, A.G. Martin, A. Ali, J.A. Cassell, Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer, BMJ Open 1 (1) (2011), e000025.
[34] D. Zhao, C. Weng, Combining PubMed knowledge and EHR data to develop a weighted Bayesian network for pancreatic cancer prediction, J. Biomed. Inform. 44 (2011) 859–868.
[35] M.W. Datta, A.M.M. Hernandez, M.J. Schlicht, A.J. Kahler, A.M. DeGueme, R. Dhir, et al., Perlecan, a candidate gene for the CAPB locus, regulates prostate cancer cell growth via the Sonic Hedgehog pathway, Mol. Cancer 5 (2006) 9.
[36] G. Napolitano, C. Fox, R. Middleton, D. Connolly, Pattern-based information extraction from pathology reports for cancer registration, Cancer Causes Control 21 (2010) 1887–1894.
[37] N.H. Heintzelman, R.J. Taylor, L. Simonsen, R. Lustig, D. Anderko, J.A. Haythornthwaite, et al., Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text, J. Am. Med. Inform. Assoc. 20 (2013) 898–905.
[38] J. Ahmed, T. Meinel, M. Dunkel, M.S. Murgueitio, R. Adams, C. Blasse, et al., CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge, Nucleic Acids Res. 39 (2011) D960–D967.
[39] K. Kadoyama, A. Kuwahara, M. Yamamori, J.B. Brown, T. Sakaeda, Y. Okuno, Hypersensitivity reactions to anticancer agents: data mining of the public version of the FDA adverse event reporting system, AERS, J. Exp. Clin. Cancer Res. 30 (2011) 93.
[40] C.-H. Lee, C.-H. Wu, H.-C. Yang, Text mining of clinical records for cancer diagnosis, in: Second International Conference on Innovative Computing, Information and Control, Kumamoto, Japan, 2007, p. 172.
[41] S.B. Park, S. Hwang, B.T. Zhang, Mining the risk types of human papillomavirus (HPV) by AdaCost, in: V. Mařík, W. Retschitzegger, O. Štěpánková (Eds.), Database and Expert Systems Applications, Springer, New York City, NY, 2003, pp. 403–412.
[42] National Cancer Institute, National Cancer Institute Factsheet, 2013 http://www.cancer.gov/cancertopics/factsheet/
[43] The Royal College of Radiologists, Standards for the Reporting and Interpretation of Imaging Investigations, 2006 http://www.rcr.ac.uk/
[44] Ö. Uzuner, Y. Juo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (2007) 550–563.
[45] L. Butt, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Classification of cancer-related death certificates using machine learning, Australas. Med. J. 6 (2013) 292–299.
[46] H. Chen, S.S. Fuller, C. Friedman, W. Hersh, Knowledge management, data mining and text mining in medical informatics, in: H. Chen, S.S. Fuller, C. Friedman, W. Hersh (Eds.), Medical Informatics: Knowledge Management and Data Mining in Biomedicine (Integrated Series in Information Systems), Springer, New York City, NY, 2010.
[47] J.J. Berman, Confidentiality issues for medical data miners, Artif. Intell. Med. 26 (2002) 25–36.
[48] European Parliament, Data Protection Directive 95/46/EC, 1995 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
[49] US Congress, Health Insurance Portability and Accountability Act, 1996 http://www.gpo.gov/fdsys/pkg/PLAW-104publ91/html/PLAW-publ91.htm
[50] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.
[51] C. Friedman, T.C. Rindflesch, M. Corn, Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine, J. Biomed. Inform. 46 (2013) 765–773.
[52] J.C. Denny, Mining electronic health records in the genomics era, PLoS Comput. Biol. 8 (2012) e1002823.
[53] International Health Terminology Standards Development Organisation, SNOMED CT, 2013 http://www.ihtsdo.org/snomed-ct/
[54] B.W. Mamlin, D.T. Heinze, C.J. McDonald, Automated extraction and normalization of findings from cancer-related free-text radiology reports, in: AMIA Annual Symposium, 2003, pp. 420–424.
[55] C.D. Bajdik, B. Kuo, S. Rusaw, S. Jones, A. Brooks-Wilson, CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes, BMC Bioinformatics 6 (2005) 78.

[56] B. Xie, Q. Ding, H. Han, D. Wu, miRCancer: a microRNA-cancer association database constructed by text mining on literature, Bioinformatics 29 (2013) 638–644.
[57] Y.-C.C. Fang, H.-C.C. Huang, H.-F.F. Juan, MeInfoText: associated gene methylation and cancer information from text mining, BMC Bioinformatics 9 (2008) 22.
[58] A. Korhonen, D.O. Séaghdha, I. Silins, L. Sun, J. Högberg, U. Stenius, Text mining for literature review and knowledge discovery in cancer risk assessment and research, PLoS ONE 7 (2012) e33427.
[59] R.S. Crowley, M. Castine, K. Mitchell, G. Chavan, T. McSherry, M. Feldman, caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research, J. Am. Med. Inform. Assoc. 17 (2010) 253–264.
[60] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, et al., BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res. 37 (2009) W170–W173.
[61] P.L. Whetzel, NCBO Team, NCBO Technology: powering semantically aware applications, J. Biomed. Semant. 4 (2013) S8.
[62] L.H. Sobin, M.K. Gospodarowicz, C. Wittekind (Eds.), TNM Classification of Malignant Tumours (UICC International Union Against Cancer), 7th ed., Wiley-Blackwell, Hoboken, NJ, 2009.
[63] S.K. Mohanty, A.L. Piccoli, L.J. Devine, A.A. Patel, G.C. William, S.B. Winters, et al., Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist, BMC Cancer 7 (2007) 144.
[64] I.A. McCowan, D.C. Moore, A.N. Nguyen, R.V. Bowman, B.E. Clarke, E.E. Duhig, et al., Collection of cancer stage data by classifying free-text medical reports, J. Am. Med. Inform. Assoc. 14 (2007) 736–745.
[65] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: American Medical Informatics Association, 2001, pp. 17–21.
[66] C. Jacquemin, Spotting and Discovering Terms Through Natural Language Processing, MIT Press, Cambridge, MA, 2001.
[67] K.B. Cohen, L. Hunter, Getting started in text mining, PLoS Comput. Biol. 4 (2008) e20.
[68] G. Schadow, C.J. McDonald, Extracting structured information from free text pathology reports, in: AMIA Annual Symposium, 2003, pp. 584–588.
[69] N. Kang, B. Singh, Z. Afzal, Mulligen EMv, J.A. Kors, Using rule-based natural language processing to improve disease normalization in biomedical text, J. Am. Med. Inform. Assoc. 20 (2013) 876–881.
[70] R. Leaman, C. Miller, G. Gonzalez, Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark, in: The 3rd International Symposium on Languages in Biology and Medicine (LBM), Jeju Island, South Korea, 2009, pp. 82–89.
[71] L. Rokach, O. Maimon, M. Averbuch, Information retrieval system for medical narrative reports, in: Flexible Query Answering Systems, Lecture Notes in Computer Science, Springer, New York City, NY, 2004, pp. 217–228.
[72] Y.-C.C. Fang, P.-T.T. Lai, H.-J.J. Dai, W.-L.L. Hsu, MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature, BMC Bioinformatics 12 (2011) 471.
[73] Y. Jin, R.T. Mcdonald, K. Lerman, A.M. Mark, M.Y. Liberman, O. Pereira, et al., Identifying and extracting malignancy types in cancer literature, in: BioLINK, Detroit, MI, USA, 2005.
[74] S.J. Nelson, K. Zeng, J. Kilbourne, T. Powell, R. Moore, Normalized names for clinical drugs: RxNorm at 6 years, J. Am. Med. Inform. Assoc. 18 (2011) 441–448.
[75] The Gene Ontology ConsortiumM. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[76] L. Hirschman, M. Colosimo, A. Morgan, A. Yeh, Overview of BioCreAtIvE task 1B: normalized gene lists, BMC Bioinformatics 6 (2005) S11.
[77] C. Rosse, J.J. Mejino, A reference ontology for biomedical informatics: the Foundational Model of Anatomy, J. Biomed. Inform. 36 (2003) 478–500.
[78] M. Gerner, G. Nenadic, C.M. Bergman, LINNAEUS: a species name identification system for biomedical literature, BMC Bioinformatics 11 (2010).
[79] I. Spasić, F. Sarafraz, J.A. Keane, G. Nenadić, Medication information extraction with linguistic pattern matching and semantic rules, J. Am. Med. Inform. Assoc. 17 (2010) 532–535.
[80] D.T. Heinze, M.L. Morsch, R.E. Sheffer Jr., M.A. Jimmink, M.A. Jennings, W.C. Morris, et al., LifeCode – a natural language processing system for medical coding and data mining, in: Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, Austin, TX, USA, 2000.
[81] P.A. Dang, M.K. Kalra, M.A. Blake, T.J. Schultz, E.F. Halpern, K.J. Dreyer, Extraction of recommendation features in radiology with natural language processing: exploratory study, Am. J. Roentgenol. 191 (2008) 313–320.
[82] Y. Yang, C.G. Chute, An application of least squares fit mapping to clinical classification, in: Annual Symposium on Computer Application in Medical Care, 1992, pp. 460–464.
[83] A. Srinivasan, The Aleph Manual, 2013 http://www.cs.ox.ac.uk/activities/machinelearning/Aleph/
[84] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (1994) 161–174.
[85] Virology NCf, HPV Sequence Database, 2013 http://ncvunledu/Angelettilab/HPV/Databasehtml
[86] J.L. Bader, M.F. Theofanos, Searching for cancer information on the internet: analyzing natural language search queries, J. Med. Internet Res. 5 (2003) e31.
[87] C. Friedman, A broad-coverage natural language processing system, in: AMIA Symposium, Los Angeles, CA, USA, 2000, pp. 270–274.
[88] N.L. Jain, C. Friedman, Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports, in: D.R. Masys (Ed.), AMIA Annual Symposium, Nashville, TN, 1997, pp. 829–833.
[89] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[90] B. Settles, ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text, Bioinformatics 21 (2005) 3191–3192.
[91] I. Spasić, M. Greenwood, A. Preece, N. Francis, G. Elwyn, FlexiTerm: a flexible term recognition method, J. Biomed. Semant. 4 (2013) 27.
[92] R. Côté, F. Reisinger, L. Martens, H. Barsnes, J.A. Vizcaino, H. Hermjakob, The Ontology Lookup Service: bigger and better, Nucleic Acids Res. 38 (2010) W155–W160.
[93] I. Spasić, S. Ananiadou, J. McNaught, A. Kumar, Text mining and ontologies in biomedicine: making sense of raw text, Brief. Bioinform. 6 (2005) 239–251.
[94] J.L. Warner, P. Anick, P. Hong, N. Xue, Natural language processing and the oncologic history: is there a match? J. Oncol. Pract. 7 (2011) e15–e19.
[95] K.S. Chan, J.B. Fowles, J.P. Weiner, Review: electronic health records and the reliability and validity of quality measures: a review of the literature, Med. Care Res. Rev. 67 (2010) 503–527.